Computational Methods and Epidemiologic Approaches for Revealing the Etiology of Autoimmune Diseases

By

Brooke Rhead

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computational Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Lisa F. Barcellos, Chair
Professor Nir Yosef
Professor Lexin Li
Professor John Colford

Summer 2019

Abstract

Computational Methods and Epidemiologic Approaches for Revealing the Etiology of Autoimmune Diseases

By

Brooke Rhead

Doctor of Philosophy in Computational Biology

University of California, Berkeley

Professor Lisa F. Barcellos, Chair

Autoimmune diseases, in which normal tissues are inappropriately attacked by the immune system, are complex diseases driven by a combination of genetic and environmental factors. Most are chronic inflammatory diseases with some treatments available but no known cures, and the disease mechanisms are not completely understood. Epigenetic factors, such as DNA methylation and microRNAs, are affected by genetic and environmental exposures and in turn affect gene expression and thus may play a role in autoimmune disease pathogenesis. In this dissertation, I employ a combination of computational, bioinformatic, statistical, and epidemiologic methods to study the role of epigenetics in autoimmune diseases in humans, and to characterize inflammatory changes in human cell lines.

Chapter one introduces some complexities of studying autoimmune diseases in humans and introduces concepts of epigenetics. Chapter two shows that naïve T cells from rheumatoid arthritis patients share DNA methylation sites with fibroblast-like synoviocytes, cells that line joints and are involved in joint inflammation. Chapter three shows that there are differences in DNA methylation in CD4+ and CD8+ T cells from multiple sclerosis patients compared to cells from healthy controls. Chapter four uses genome-wide association study results to implicate specific microRNAs and tissues in pediatric-onset multiple sclerosis. Chapter five shows that the inflammatory cytokine tumor necrosis factor alpha drives DNA methylation and transcriptional changes and activates autoimmune disease genes in endothelial cells. Chapter six is a summary of conclusions and key findings.

For my mom and dad.

Acknowledgments

Thank you first to my excellent and capable advisor, Lisa Barcellos. Her guidance has been invaluable in my development as a scientist. Her support was unwavering, even when we disagreed. Thank you to my dissertation committee members, Nir Yosef, Lexin Li, and Jack Colford, for being available and approachable, for offering good advice, and for making my qualifying exam surprisingly non-terrible. Thank you to my many collaborators: Lindsey Criswell and Emmanuelle Waubant at UCSF, Cathy Schaefer at Kaiser Permanente, Hanne Harbo and Steffan Bos at the University of Oslo, Jeannette Lechner-Scott at the University of Newcastle, and the members of the U.S. Pediatric M.S. Network, for sharing not only your data, but your thoughtful input, edits, and support. Thank you to Kate Chase and Xuan Quach for keeping the Comp Bio program running smoothly, Hong and Diana Quach for keeping the lab running smoothly, and Indro Federigo for keeping the servers running smoothly, always with good nature and aplomb.

Thank you to my co-first-authors: to Calliope Hollingue, for supplying unmitigated enthusiasm and planning capability, and to Ina Brorson, for supplying the idea that we will figure everything out and a true appreciation of Norway and aquavit. Thank you to my 2013 cohort, the first class of UC Berkeley Comp Bio PhD students: Rob Tunney, Jim Kaminski, Jeff Spence, and Amy Ko, and to our honorary cohort members from 2014: Shaked Afik and David DeTomaso. We bonded, found labs, and made it to graduation. Thank you to my past and present labmates: Xiaorong Shao, Milena Gianfrancesco, Michael Cole, Farren Briggs, Giovanna Cruz, Amanda Mok, Olivia Solomon, Calvin Chi, Cam Adams, and Mary Horton, for being brilliant, steady voices of reason, and for enabling much fun at conferences. Thank you to Bryn Reinstadler and Jiyng Zou for your outstanding work with me on your master's and undergraduate projects. Thank you to Dr. Erika Garcia for keeping me grounded and for making me laugh like no one else. Thank you to my neighbors/colleagues Marlisa Pillsbury and Jason Huff for letting me know my first semester here that it would get better (it did!). I tell people that getting a PhD is worth it, if, for nothing else, that you meet amazing people. You all are the reason why.

Thank you to my many friends at the UCSC Genome Browser for propelling me here and for celebrating me when I got here, especially to Ann Zweig, Bob Kuhn, Donna Karolchick, Jim Kent, David Haussler, Melissa Cline, Jorge Garcia, Jeltje van Baren, Jonathan Casper, Brian Rainey, Max Haeussler, Brian Lee, Matthew Spier, Mary Goldman and Karen Miga. Thank you to Mark Diekhans for getting me into this mess in the first place. Thanks to Pauline Fujita and Mike Loh for telling me I could absolutely get a PhD and questioning the wisdom of doing so. Thank you to Kayla Gross (née Smith) for thinking it was a bad idea. Thank you to Liz Beacham, Jocelyn McDaniel, and Nannette Nelson for visiting me and cheering me on. Thank you to my family, for listening to me while I despaired, complained, and ultimately succeeded. Thank you to my Uncle Jacques for mapping this course a long time ago and for checking in on me along the way. Thank you to my partner Geordie Burdick for making the last couple of years a whole lot more fun.

Finally, thank you to the National Institutes of Health (and the taxpayers) for funding me.

Table of Contents

Chapter 1 - Introduction

Complex human diseases—those caused by a combination of genetic variants and environmental exposures—are difficult to study for a variety of reasons. First, these exposures are not experienced in isolation, and for some exposures, the size of their effects may depend on the presence of other exposures. For example, cigarette smoking increases the risk of developing multiple sclerosis, but the risk is compounded in those with a genetic variant in an enzyme that metabolizes products from tobacco smoke.(1, 2) Identifying and disentangling multiple potential causes of human disease can be complicated because, with the exception of randomized controlled trials to evaluate specific interventions, exposures cannot be studied experimentally in humans and we must rely instead on observational data, which can be plagued by confounding factors and biases. Another reason complex diseases are difficult to study is that many genetic exposures are either common but increase the risk of disease by only a small amount, or they increase risk substantially but are very rare; in either case large studies are required to obtain the statistical power needed to detect their effects. Happily, large-scale genome-wide association studies (GWASs) have been immensely successful in the past decade in identifying genetic variants that predispose individuals to disease by even very small amounts. Many argue that the limits of this line of study have been reached and that we are now in a "post-GWAS era," in which the major research goals have shifted to better understanding the function of disease-associated genetic variants and to looking at orthogonal types of data, such as epigenetic data, for clues about how genetic and environmental risk factors actually cause disease.(3–6) This dissertation focuses on post-GWAS era research problems. A final layer of complexity for studying these types of problems is that, although all cells contain the same genetic variants (with the exception of somatic mutations), different genes are active in different tissue types, and the processes that contribute to disease can occur in any organ or tissue. Therefore, tissue type is a major consideration when investigating epigenetic disease associations.

Because of these complexities, studying the etiology of complex diseases requires utilizing methods from multiple fields. Computational and bioinformatics methods are needed to handle the staggering variety and amount of data that is generated in genetic studies; sophisticated statistical methods are needed to find diminishingly small signals in a large amount of data and noise; and sound epidemiologic methods are needed to design studies that reduce as much potential bias and confounding as possible, and to analyze imperfect observational data to the best of our ability. The overlapping goals and methods of these fields have converged in the field of genetic epidemiology. My goal as a genetic epidemiologist is to utilize these tools to better understand the predictors, pathogenesis, and prognosis of autoimmune diseases, and to help identify potential new therapeutic targets.

Autoimmune diseases are chronic, often inflammatory, diseases in which the immune system mistakenly attacks normal tissues. Over 80 autoimmune diseases have been identified, and they affect an estimated 4.5% of the population worldwide, with a higher prevalence in women than men.(7, 8) My work focused primarily on rheumatoid arthritis

(RA), in which joint tissues are inflamed and damage, and multiple sclerosis (MS), which causes inflammation and damage in the central nervous system (CNS). I investigated the epigenetic causes of both RA (chapter 2) and MS (chapters 3 and 4).

Epigenetics refers to heritable differences in gene expression that are not due to differences in the DNA sequence. One of the most well studied epigenetic features is DNA methylation—the addition of a methyl group to cytosines in DNA. Initially, DNA methylation was understood mainly as a mechanism to silence gene expression, but it is now appreciated as a dynamic regulator that can either increase or decrease expression, or determine which splice isoform of a gene is expressed, depending on where it is located.(9, 10) Like the DNA sequence itself, DNA methylation patterns are inherited. However, unlike the DNA sequence, DNA methylation is influenced by changes to the environment. Age, sex, smoking, air pollution, diet, exercise, stress, and medications have all been associated with differences in DNA methylation.(11–15) DNA methylation patterns are tissue-specific, which makes them both interesting and challenging to study, as they can inform us about differences that exist in disease-relevant tissues (joint tissues in RA or CNS tissues in MS, for example), but are difficult to obtain, since extracting biological samples from most tissue types other than saliva or blood is often too invasive to do on a large scale.

MicroRNAs (miRNAs) are another gene regulatory mechanism and can be considered part of the epigenetic machinery.(16, 17) miRNAs are short, ~22 nucleotide non-coding RNAs that down-regulate gene expression by binding to complementary sequence on messenger RNAs and target them for degradation, preventing them from being translated into proteins. Any given miRNA can target multiple genes, and genes can be targeted by multiple miRNAs, thus miRNAs have the potential to target entire networks of genes at once. As with DNA methylation, tissue type is important to consider when investigating miRNAs, since different miRNAs are present in different tissues.

Chapter 2 is an investigation of whether DNA methylation patterns found in synoviocytes (a type of cell found in joints) from RA patients can also be detected in the immune cells circulating in blood of RA patients.

Chapter 3 is a comparison of DNA methylation in T cells of MS cases and controls. T cells are the immune cell type responsible for attacking CNS tissue in MS.

Chapter 4 is an examination of miRNA contributions to pediatric-onset MS that can be inferred from GWAS data.

Chapter 5 is a characterization of the gene expression and DNA methylation changes that occur when tumor necrosis factor alpha (TNFa), an inflammatory cytokine, is increased in human endothelial cells. Endothelial cells line the walls of blood vessels and are important sites of inflammation in infectious and inflammatory disease. This chapter is unique in that it is an experimental study conducted on cell lines rather than an observational study.

This body of work demonstrates the utility of applying epidemiologic approaches and computational methods to the study of autoimmune disease pathogenesis, using both observational and experimental studies, with the goals of better understanding disease mechanisms and uncovering future avenues of research for disease treatment and prevention.

References

1. Briggs FBS, et al. (2014) Smoking and Risk of Multiple Sclerosis. *Epidemiology* 25(4):605–614.
2. Olsson T, Barcellos LF, Alfredsson L (2016) Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. *Nat Rev Neurol* 13(1):26–36.
3. Huang Q (2015) Genetic Study of Complex Diseases in the Post-GWAS Era. *J Genet Genomics* 42(3):87–98.
4. Gallagher MD, Chen-Plotkin AS (2018) The Post-GWAS Era: From Association to Function. *Am J Hum Genet* 102(5):717–730.
5. Baranzini SE (2018) The era of GWAS is over – Commentary. *Mult Scler J* 24(3):260–261.
6. Wijmenga C, Zhernakova A (2018) The importance of cohort studies in the post-GWAS era. *Nat Genet* 50(3):322–328.
7. Hayter SM, Cook MC (2012) Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. *Autoimmun Rev* 11(10):754–765.
8. Ngo S, Steyn F, McCombe P (2014) Gender differences in autoimmune disease. *Front Neuroendocrinol* 35:347–369.
9. Tirado-Magallanes R, Rebbani K, Lim R, Pradhan S, Benoukraf T (2017) Whole genome DNA methylation: beyond genes silencing. *Oncotarget* 8(3):5629–5637.
10. Luo C, Hajkova P, Ecker JR (2018) Dynamic DNA methylation: In the right place at the right time. *Science (80-)* 361(6409):1336–1340.
11. Horvath S (2013) DNA methylation age of human tissues and cell types. *Genome Biol* 14(10):R115.
12. Martin EM, Fry RC (2018) Environmental Influences on the Epigenome: Exposure-Associated DNA Methylation in Human Populations. *Ssrn*. doi:10.1146/annurev-publhealth-040617-014629.
13. Ronn T, Ling C (2014) The impact of exercise on DNA methylation of genes associated With type 2 diabetes and obesity in human adipose tissue. *US Endocrinol* 10(1):64–66.
14. Vinkers CH, et al. (2015) Traumatic stress and human DNA methylation: a critical review. *Epigenomics* 7(4):593–608.
15. Lee SW, et al. (2018) Whole-genome methylation profiling of peripheral blood mononuclear cell for acute exacerbations of chronic obstructive pulmonary disease treated with corticosteroid. *Pharmacogenet Genomics* 28(3):78–85.
16. Vasilatou D, Papageorgiou SG, Dimitriadis G, Pappa V (2013) Epigenetic alterations and microRNAs: New players in the pathogenesis of myelodysplastic syndromes. *Epigenetics* 8(6). doi:10.4161/epi.24897.
17. Koch MW, Metz LM, Kovalchuk O (2013) Epigenetics and miRNAs in the diagnosis and treatment of multiple sclerosis. *Trends Mol Med* 19(1):23–30.

Chapter 2 - Rheumatoid arthritis naïve T cells share hypermethylation sites with synoviocytes

Abstract

Objective: Our study aimed to determine whether differentially methylated CpGs in synovium-derived fibroblast-like synoviocytes (FLS) of rheumatoid arthritis (RA) patients were also differentially methylated in peripheral blood samples.

Methods: We measured 371 genome-wide DNA methylation profiles from 63 RA cases and 31 controls, in CD14+ monocytes, CD19+ B cells, CD4+ memory T cells and CD4+ naïve T cells, using Illumina HumanMethylation450 (450k) BeadChips.

Results: We found that of 5,532 hypermethylated FLS candidate CpGs, 1,056 were hypermethylated in CD4+ naïve T cells of RA cases compared to controls. Using a second set of CpG candidates based on SNPs from a genome-wide association study (GWAS) of RA, we found one significantly hypermethylated CpG in CD4+ memory T cells and 18 significant (6 hypomethylated, 12 hypermethylated) CpGs in CD4+ naïve T cells. A prediction score based on the hypermethylated FLS candidates had an area under the curve (AUC) of 0.73 associated with RA case status, which compared favorably to the association of RA with the *HLA-DRB1* shared epitope (SE) risk allele and with a validated RA genetic risk score.

Conclusion: FLS-representative DNA methylation signatures derived from blood may prove to be valuable biomarkers for RA risk or disease status.

Introduction

RA is a chronic inflammatory disease with the potential to cause substantial disability, primarily due to the erosive and deforming process in joints. It is the most common systemic autoimmune disease, with a worldwide prevalence approaching 1% (1,2). RA etiology is complex, with both genetic and non-genetic contributions. A rigorous assessment of RA heritability using twin studies suggests that 50-60% of the occurrence of RA in twins is explained by genetic effects (3). Approximately 50% of this genetic contribution can be explained by genes in the major histocompatibility complex (MHC) (3). In addition, at least 101 independent non-MHC risk loci have been identified (4). A role for environmental factors is also supported, but currently exposure to tobacco smoke is the only well-established risk factor (5).

DNA methylation is an epigenetic modification resulting from the addition of a methyl group to a cytosine base at positions in the DNA sequence where a cytosine is followed by a guanine (“CpGs”), which can lead to altered expression of DNA. DNA methylation is essential for proper mammalian development and other functions, and methylation patterns are affected by environmental changes. Methylation status is also influenced by the interaction between genetics and environment, and a growing number of human

diseases have been associated with aberrant DNA methylation. (6) Maintenance of DNA methylation is critical for the development and function of immune cells (6,7).

Altered patterns of DNA methylation at CpG sites have been observed in individuals with RA. A 1990 study by Richardson et al. found that global methylation of genomic DNA from T cells of RA patients was lower when compared to T cells of healthy controls (8). Altered methylation patterns have also been observed in small studies of specific genes in RA, including the promoter regions of *IL6* using peripheral blood mononuclear cells (PBMCs) and *DR3* (alternative name *TNFRSF25*) using synovial fibroblasts (9,10). Liu et al. studied global DNA methylation among 129 Taiwanese individuals and found that RA patients were characterized by significantly lower levels of DNA methylation in PBMCs compared to controls (11). Recently, Glossop et al. identified about 2,000 differentially methylated CpGs in both T- and B-lymphocytes between treatment-naïve patients with early RA and healthy individuals, and in a separate analysis found that DNA methylation profiles from synovial fluid-derived FLS had similarities with the profiles from tissue-derived FLS (12,13).

A recent investigation identified 15,220 differentially methylated CpG sites in synovium-derived fibroblast-like synoviocytes (FLS) between RA patients and either osteoarthritis or normal controls that appear to distinguish RA cases from non-RA controls (Whitaker et al. (14) and personal communication). These 15,220 FLS CpGs are the candidate sites for the current investigation. FLS in the synovial intimal lining of joints have key roles in the production of cytokines that perpetuate inflammation, and the production of proteases that contribute to cartilage destruction in RA (15). An overlap in the methylation pattern between FLS and peripheral blood cells could be indicative of disease-associated biological processes detectable in the periphery. Because peripheral blood is easily accessible, such signatures may be useful biomarkers for RA risk or disease status.

Materials and Methods

Study Design

Participants included 63 female RA cases (18 of age or older and met the 1987 American College of Rheumatology criteria for RA (16,17)) and 31 female unaffected controls (locally based), all of European ancestry. Table 1 summarizes characteristics of our study population. All participants provided a peripheral blood sample for genotyping and measurement of methylation.

Genotyping

Study participants were genotyped using Illumina HumanOmniExpress, HumanOmni-ExpressExome, or Human660W-Quad Beadchips, which were read on an Illumina HiScan array scanner. Genotype results were merged using PLINK v1.07 (18), and only SNPs assessed by all three chips were retained for analysis. SNPs with failed genotype calls in 10% or more of individuals, with a minor allele frequency of less than 1%, or found

to not be in Hardy-Weinberg equilibrium ($p \leq 0.000001$) in controls were removed from analysis.

Ancestry

EIGENSTRAT (19) was used to visualize ancestral clustering of the study population relative to individuals from 11 HapMap populations (20). Self-identified individuals of European ancestry clustered with Utah residents with ancestry from northern and western Europe/Tuscans in Italy (CEPH/TSI) as expected. We excluded self-reported individuals not of European ancestry because of the potential for confounding. Figure S3 shows the ancestral clustering of our final sample of self-identified European-ancestry participants.

Cell Sorting

Whole blood was collected in four 10ml EDTA collection tubes from each subject. PBMCs were isolated using Ficoll-Paque density gradient and stained with conjugated monoclonal antibodies against CD45 FITC, CD19 PE, CD45RA PE-CY7 (all BD Pharmingen), CD3 Brilliant Violet 421, CD4 CF594 (both BD Horizon), CD14 APC (BD Biosciences) and CD27 APC-eFlour780 (ebioscience). Cells were then stored overnight in buffer at 4°C and sorted the following day, on a BD FACSAria cell sorter (BD Biosciences). The following populations were gated for sorting following exclusion of debris and doublets: monocytes (CD45+CD14+); B cells (CD45+CD14-CD3-CD19+); naïve CD4+ T cells (CD45+CD14-CD19-CD3+CD4+CD27+CD45RA+) and memory CD4+T cells (CD45+CD14-CD19-CD3+CD4+CD45RA-). Cell counts and purity checks were performed after sorting, and then cells were stored frozen as a pellet at -80°C.

Validation of overnight cell storage

To enable DNA methylation profiling of a large number of FACS samples, a protocol for storing blood samples overnight prior to sorting was established and validated. Whole blood was collected in ten 10ml EDTA collection tubes from a single individual. PBMCs were isolated and stained as described above, and then either sorted the same day or stored overnight in buffer at 4°C and sorted the following day. Paired DNA samples from the two time points were collected from all four cell types. All DNA samples were quantified using a Nanodrop spectrophotometer. All samples underwent bisulfite conversion on the same day and were assayed on Illumina 450k BeadChips simultaneously.

Methylation

A total of 371 genome-wide DNA methylation profiles were generated using the Illumina Infinium HumanMethylation450 BeadChip kit and read on an Illumina HiScan array scanner. A β value, the ratio of the methylated probe intensity to the overall (methylated plus unmethylated) intensity, was derived for each CpG site. We performed an extensive QC process: Illumina GenomeStudio software was used to examine Jurkat controls,

between chip/within chip variation, and replicate samples. All replicate samples had r^2 values greater than 0.99 and Jurkat replicates showed r^2 greater than 0.98. Background signal was subtracted using the methylumi R package "noob" method (21) and samples were normalized with All Sample Mean Normalization (ASMN) (22) followed by beta-mixture quantile normalization (BMIQ) (23) to correct for type I and type II probe differences. Multidimensional scaling (MDS) plots for each cell type before and after background subtraction and normalization were examined to assess for the presence of batch effects. Batch effects were found to be minimal, and were reduced following data normalization (see Figures S4 and S5 for an example). 286 CpG sites with low detection rates (read $p>0.05$) in more than 20% of samples were removed from analysis, and one sample with low detection rates (read $p>0.05$) in more than 20% of sites. The following CpG sites were also removed from analysis: the 65 non-CpG "rs" SNP probes included in the 450k BeadChip, 30,969 sites with probes predicted to hybridize to more than one location in the genome after bisulfite conversion ("cross-reactive probes") identified by Chen et al., and 28,355 sites with a known polymorphism at the site being measured ("polymorphic CpGs") identified by Chen et al. that were either present in our European-ancestry population or present in Europeans in the 1,000 Genomes Project (24). The final data set used for analysis consisted of 428,232 CpG sites in 371 samples (94 CD14+ monocyte samples, 91 CD19+ B cell samples, 94 CD4+ memory T cell samples, and 92 CD4+ naïve T cell samples).

An MDS plot of all 371 samples (Figure 3) shows that each of the four immune cell types cluster together as expected based on their DNA methylation patterns. Differences in methylation among different cell types are much larger than the differences between cases and controls within each cell type, as expected. There is greater scattering for B cells, which is reflective of the diversity of that cell type, versus monocytes and the T cell subpopulations examined in this study.

Wilcoxon Rank Sum Tests

Four immune cell types were assayed for each individual: CD14+ monocytes, CD19+ B cells, CD4+ memory T cells, and CD4+ naïve T cells. DNA hypermethylation or hypomethylation in RA cases relative to controls consistent with methylation differences seen in FLS was evaluated separately for each immune cell type. For each of the hypermethylated (n=5,532) and hypomethylated (n=8,406) candidate CpGs from the FLS study, we used a one-tailed Wilcoxon rank sum test to assess differences in the median β value between RA cases and controls. P-values were adjusted using the Benjamini-Hochberg method for controlling the false discovery rate (25). We controlled the error rate for 5,532 or 8,406 tests, depending on the candidate list. Methylation changes at a second set of 1,788 candidate CpG sites in 98 genes deemed likely to be important to RA biology based on a recent genome-wide association study (GWAS) meta-analysis of >100,000 subjects (4) were also evaluated, and an exploratory association analysis was conducted using all CpGs on the 450k BeadChip. For the GWAS candidate CpGs and the chip-wide tests, we used a two-tailed Wilcoxon rank sum test, controlling for 1,676 tests and 428,232 tests, respectively.

ReFACTor Principal Component

In order to determine whether cell subtype proportions in the sorted cells were confounding results, we performed Reference-Free Adjustment for Cell-Type composition (26) (ReFACTor), in which principal component (PC) analysis is performed on a subset of sites that are informative with respect to the cell composition in the data. ReFACTor finds the most informative sites in an unsupervised manner. To measure the potential confounding, we examined quantile-quantile (QQ) plots for each cell type for a standard epigenome-wide association study (EWAS), using only the methylation sites with a mean methylation level in the range of 0.2-0.8, following a suggestion of Liu et al. to remove consistently methylated and consistently unmethylated probes when performing EWAS (27). Deflation was observed in the QQ-plots of all cell types except CD4+ naïve T cells, implying deficient power. To assess the expected QQ-plot under the condition of power deficiency, we permuted the phenotype and repeated the EWAS analysis, and repeated this procedure 100 times for each cell type. To determine whether a correction was required in the cell types, we used the genomic control lambda measurement of inflation (28). We considered the median lambda of the 100 EWAS executions as the expected lambda. The approach was to add ReFACTor PCs to the analysis until the inflation was corrected with respect to the expected lambda (29). Only the CD4+ naïve cells were found to be inflated, and adjusting for the first ReFACTor component removed this inflation, suggesting possible cell substructure in the CD4+ naïve cells. ReFACTor was executed on the CD4+ naïve T cell data with parameter K=2. We added the first PC (PC1) in logistic regression models to evaluate results that are adjusted for confounding by cell substructure.

Logistic Regression Models

To evaluate possible confounding effects, logistic regression models of RA case status were carried out against each FLS CpG that was significant at $q<0.05$ in the Wilcoxon tests (1,056 models), adjusting for smoking, age, batch (date the plate was run), and PC1 calculated from the ReFACTor analysis described above, which aims to quantify cell substructure (26). Unadjusted models were compared to models adjusted for age only; ever having smoked only; batch only; ReFACTor PC1 only; age, smoking and batch combined; and age, smoking, batch and ReFACTor PC1 combined.

ROC curve analyses

Receiver operator characteristic (ROC) curve analysis was used to explore the potential for the FLS sites to serve as a biomarker for the RA disease process, compared to the potential of a validated genetic risk score for RA (30,31) and the presence or absence of *HLA-DRB1* shared epitope alleles (32,33). The hypermethylation score for each person was calculated by summing the beta values across the 1,056 FLS significantly differentially methylated loci. A continuous weighted genetic risk score was also calculated, based on the publications by Yarwood et al. (31) and Eyre et al. (30) The

genetic risk score included 43 of the 45 non-HLA SNPs (rs13397 and rs59466457 were missing), and it was calculated by multiplying the number of copies of risk alleles, using probability data from genome-wide imputation, for each SNP by the natural logarithm of the odds ratio as reported in Eyre et al. (30), and summing these values across the 43 SNPs for each person. Presence of the shared epitope was coded as a binary variable. Individuals with one or more copies of the following alleles were assigned a value of one for the shared epitope: *HLA-DRB1**0101, *0102, *0401, *0404, *0405, *0408, or *1001 (34). The pROC package in R was used to plot each of these variables as a predictor with RA case-status as the response variable (35).

To determine the influence of adjusting for potential confounders of the hypermethylation score, we created two additional hypermethylation scores: the first based on the 830 FLS sites that remained significant ($p<0.05$) in the logistic regression models after adjusting for age, smoking, and batch, and the second based on the 79 FLS sites that remained significant ($p<0.05$) after adjusting for age, smoking, batch and ReFACTor PC1.

Study Approval

Written informed consent was received from all participants prior to inclusion in this study, and research was in compliance with the Helsinki Declaration. Institutional Review Board approval was in place at UC San Francisco where study subjects were recruited.

Results

Validation of overnight cell storage

Methylation profiles for isolated cell populations were not impacted by overnight storage (correlation between profiles derived from all paired samples was very high ($r^2>0.997$)). Details are summarized in Supplementary Text 1.

Candidate FLS CpG results

After adjusting p-values from the Wilcoxon rank sum tests for multiple testing by controlling the false discovery rate (FDR; p-values adjusted for multiple testing hereby referred to as q-values), 1,056 significantly hypermethylated CpG sites in CD4+ naïve T cells had $q<0.05$ (Table S1). There were no significant sites at this threshold for the hypomethylated candidates in CD4+ naïve T cells, nor in any of the remaining cell types (CD14+ monocytes, CD19+ B cells and CD4+ memory T cells), for either the hyper- or hypomethylated candidates. Results are summarized in Table 2.

Logistic Regression Results

Logistic regression analysis was conducted with RA case status as the outcome and methylation beta value as the predictor variable for each of the 1,056 FLS CpG. 1,035 CpGs were significant ($p<0.05$, one-sided) in the unadjusted model, 830 remained

significant when adjusting for age, smoking and batch together, and 79 remained significant when adjusting for age, smoking, batch, and ReFACTor PC1. Results are summarized in Table S2, and the shifts in p-values with different models are visualized in Figure 2.

Comparison of Methylation Profiles to Shared Epitope and Genetic Risk Score

The association of hypermethylation in CD4+ naive T cells with RA was compared to a weighted genetic risk score for non-HLA risk alleles, and presence or absence of the *HLA-DRB1* shared epitope, a major genetic risk factor for RA (36). The hypermethylation score and shared epitope models performed similarly. Figure 1 shows the three ROC curves and Table 3 summarizes the point estimates and 95% confidence intervals for the area under the curve (AUC) for each model. The hypermethylation score had the largest AUC of 72% (61%-83%). The shared epitope had an AUC of 66% (56%-76%), and the genetic risk score had an AUC of 51% (38%-63%). The AUC for the hypermethylation score based on the 830 CpGs significant at $p<0.05$ after adjusting for age, smoking, and batch in the logistic regression models was 71.8% (61.0%-82.7%), which is similar to the hypermethylation score using unadjusted CpGs significant after the Wilcoxon test. The AUC using only the 79 CpGs significant after adjusting for age, smoking, batch, and ReFACTor PC1 was 80.7% (71.3%-90.1%). Results are summarized in Table 3.

Candidate GWAS CpG results

For this set of Wilcoxon rank sum tests (1,676 CpGs), one CpG (hypermethylated) in CD4+ memory cells and 18 CpGs (6 hypomethylated, 12 hypermethylated) in CD4+ naïve T cells were significantly associated ($q<0.05$) with RA susceptibility. Results are summarized in Table S4. We also carried out logistic regression analysis using RA status as outcome for each of the 18 CpGs that were differentially methylated in CD4+ naïve T cells, adjusting for various covariates. Results are summarized in Table S5.

Genome-wide results

Results of the genome-wide tests of differences in methylation are summarized in Table S6. No CpG sites were significantly differentially methylated after multiple testing correction (adjusting the p-value for 428,232 tests). Differences in global methylation were investigated by comparing mean methylation levels in cases and controls (Table S7). No significant differences were observed for any cell type.

Discussion

In the current study, hypermethylated CpG sites previously identified in FLS of RA cases relative to osteoarthritis or healthy controls were also distinguished in CD4+ naïve T cells from peripheral blood of RA cases relative to healthy controls. Our results show a disease-associated signature can be observed in cells obtained from whole blood, which is more accessible for clinical or epidemiologic studies compared to synovial fluid.

Our work extends recent findings demonstrating DNA methylation profiles in peripheral blood mononuclear cells differ between RA cases and controls (12). While Glossop et al. observed differences in both B-lymphocytes and T-lymphocytes, most results from the current study were confined to CD4+ naïve T cells. However, taken together, the combined findings increase the evidence that peripheral blood cells contain a DNA methylation signature that can distinguish RA cases from controls. Furthermore, the identification of DNA methylation profile differences in T cells detected in treatment naïve patients by Glossop suggests there are methylation changes important in RA that are not a consequence of medication or long disease duration.

The 1,056 differentially methylated candidate FLS CpGs associated with RA in this study were limited to the CD4+ naïve T cell population. Most of the observed differences were small, with a difference in median β value of less than 10% between RA cases and controls. Of the 1,056 sites, 517 had a methylation difference of greater than 1% (Table S1). These 517 sites resided in 357 genes as well as intergenic regions, and across all chromosomes. It is uncertain what effect size is biologically meaningful for DNA methylation. Some researchers impose a threshold of 5% or 10% difference in methylation to consider results relevant (37), while others include modest effect sizes (38). One recent study showed replicable methylation differences associated with smoking ranging from 1.2% to 24% (39). Though differences in this study were small, they were robust, surviving stringent multiple testing correction. A hypermethylation score constructed from the significant 1,056 sites predicted RA case status with an AUC of 73%, and awaits validation in an independent dataset. The hypermethylation score based on the 830 CpG sites with $p<0.05$ after adjusting for smoking, age, and batch in the logistic regression models had a similar AUC of 71.8%, suggesting the score was not strongly influenced by these covariates. The hypermethylation score calculated using the 79 CpG sites with $p<0.05$ after adjusting for smoking, age, batch and ReFACTor PC1 in the logistic regression models had a slightly higher AUC of 80.7%, suggesting that adjustment for possible cell substructure may improve the ability of our FLS CpG sites score to serve as a biomarker for RA. Because DNA methylation was measured subsequent to RA diagnosis, we cannot tell with certainty whether the FLS methylation signature in the CD4+ naïve T cells predicts RA diagnosis or is a biomarker of the disease process.

One of the top 10 (most significant p-value) CD4+ naïve T cell replicated sites, cg21480173, was found in the gene *TYK2*, which has been associated with RA and other autoimmune diseases (40). The remaining 9 top hits were found in the following genes: *PRKAR1B, ABCC4, COMT, CAI2, MCF2L, GALNT9, C7orf50*, or non-gene regions, which have not been previously associated with RA. Results demonstrate that novel genes related to RA may be discovered through DNA methylation analysis. We also observed differential methylation in CpG sites that reside in genes that have previously been associated with RA (4). For example, two of the CpGs reside in the promoter regions for both *GATA3* and *GATA3-AS1* (cg17566118 and cg15852223), and both are hypomethylated in RA cases relative to controls. It is important to note our results were not due to genetic variation or genetic ancestry differences between cases and controls.

The lack of significant findings in cell types other than CD4+ naïve T cells suggests that CD4+ naïve T cells are particularly relevant to RA through epigenetic mechanisms involving DNA methylation. There is strong evidence from previous studies that aberrant T-cell activation pathways are involved in the pathogenesis of RA, including in the naïve T cell population, which have not yet participated in immune responses (41). CD4+ naïve T cells from RA patients have been shown to have premature senescence; to be defective in up-regulating telomerase due to deficiencies in telomerase component human telomerase reverse transcriptase (hTERT); to have increased DNA damage load and apoptosis rates; to not metabolize equal amounts of glucose as healthy control cells of the same age; and to generate less ATP (42–45). While our methylation findings need to be replicated, the striking CD4+ naïve T cell results, and the existing literature on abnormalities in this cell population in RA, suggest that the methylation changes we observed may be involved in disease pathogenesis. However, it is also plausible that methylation changes are a response to the disease process itself or a result of exposure to medications. Additional studies involving patients with early or pre-clinical disease will be required to determine when in the course of the disease process such differential methylation patterns occur. Longitudinal studies may also help elucidate why results from the current study support a hypermethylation signature in RA, in contrast with hypomethylation which has been demonstrated in previous studies (8,11). Hypermethylation may occur at a specific point along the course of RA, or may be specific to the FLS-associated sites rather than the global methylome.

Results from logistic regression modeling suggest that although some variables are confounding the relationship between methylation and RA case status, evidence for association persists. Specifically, adjusting for age or smoking did not markedly impact the number of FLS CpGs that were significantly associated with RA at $p<0.05$. Adjusting for batch or ReFACTor PC1 reduced the number of statistically significant CpGs by ~ 200, but many remained statistically significant (841 adjusting for batch, 837 adjusting for ReFACTor PC1). Even when controlling for all four of these variables, 79 CpGs remained significant. Figure 2 visually represents the shifting of p-values across these regression models. Evidence for association also persisted in analysis of GWA candidates, even in fully adjusted models (Table S5).

Strengths and Limitations

This study has many strengths. DNA methylation profiles were analyzed in four sorted cell types for 94 individuals who are all females of European ancestry, which reduced the genetic heterogeneity of the study population. Examination of individual cell types from FACS-sorted blood allowed us to measure methylation results with more confidence, rather than relying on whole blood and cell type proportions (46). Restriction of the study to females eliminates the possibility of confounding by sex. Also, since RA affects women at a 3:1 ratio relative to men, results are generalizable to the group that experiences the greatest disease burden.

Stringent quality control of the methylation data, as described in the methods, is another strength. In addition to standard QC steps of background subtraction, normalization, and removal of sites with low quality scores, CpG sites with known SNPs in individuals of European ancestry at the cytosine or guanine being measured on the 450k BeadChip were removed, which is important because methylation measurements for CpG sites harboring SNPs are likely to simply reflect genetic polymorphism at that site rather than truly measuring methylation. We also removed from analysis CpG sites with cross-reactive sequencing probes on the 450k BeadChip, i.e., probes that could hybridize to more than one location in the genome and reflect methylation at two different genomic locations rather than only the intended target site. Rigorous quality control measures increase confidence that the observed differential methylation is an accurate reflection of the disease biology and not due to artifacts.

Both whole genome and whole methylome data were utilized in the current study. The whole genome data allowed us to determine genetic ancestry for all participants. The original FLS study by Whitaker et al. involved anonymous samples, and the authors did not have ethnicity or race information (14). Therefore, it is possible that we are underestimating the overlap between FLS and CD4+ naïve sites if we are comparing different ethnicities in the CD4+ naïve T cell and FLS group. Lastly, we were able to demonstrate that even after controlling for age, smoking, batch and possible cell substructure (ReFACTor PC1), a number of FLS and GWA candidate sites remain significantly associated with RA.

This study also has limitations. We could not assess temporality between methylation and case-status. Results may be confounded by case-specific factors such as medication and inflammation. Indeed, other studies have observed associations between methylation and medications (47,48); however, the case-control nature of the current study did not allow us to adjust for effects of RA medications since they were present only among RA cases.

Our findings are restricted to CpG sites that are represented on the 450k BeadChip. The BeadChip prioritized inclusion of features such as RefSeq genes; CpG Islands, shores and shelves; areas of the genome such as the MHC region; and sites known to be in important to cancer (49,50). Therefore, additional CpG sites relevant to RA may be missing. Further, although our ROC analysis demonstrates that differential methylation of about 1,000 CpGs in peripheral blood has the potential to distinguish RA cases from controls, our hypermethylation score needs to be tested as a predictor in an independent data set.

References

1. Cojocaru M, Cojocaru I, Silosi I, Vrabie C, Tanasescu R. Extra-articular manifestations in rheumatoid arthritis. *Maedica (Buchar)* 2010;5:286–91.

2. Gabriel SE, Michaud K. Epidemiological studies in incidence, prevalence, mortality, and comorbidity of the rheumatic diseases. *Arthritis Res Ther* 2009;11:229.

3. MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K, et al. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum* 2000;43:30–37.

4. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 2014;506:376–81.

5. Källberg H, Ding B, Padyukov L, Bengtsson C, Rönnelid J, Klareskog L, et al. Smoking is a major preventable risk factor for rheumatoid arthritis: estimations of risks after various exposures to cigarette smoke. *Ann Rheum Dis* 2011;70:508–511.

6. Robertson KD. DNA methylation and human disease. *Nat Rev Genet* 2005;6:597–610.

7. Ohkura N, Kitagawa Y, Sakaguchi S. Development and Maintenance of Regulatory T cells. *Immunity* 2013;38:414–423.

8. Richardson B, Scheinbart L, Strahler J, Gross L, Hanash S, Johnson M. Evidence for impaired T cell DNA methylation in systemic lupus erythematosus and rheumatoid arthritis. *Arthritis Rheum* 1990;33:1665–1673.

9. Takami N, Osawa K, Miura Y, Komai K, Taniguchi M, Shiraishi M, et al. Hypermethylated promoter region of DR3, the death receptor 3 gene, in rheumatoid arthritis synovial cells. *Arthritis Rheum* 2006;54:779–787.

10. Nile CJ, Read RC, Akil M, Duff GW, Wilson AG. Methylation status of a single CpG site in the IL6 promoter is related to IL6 messenger RNA levels and rheumatoid arthritis. *Arthritis Rheum* 2008;58:2686–2693.

11. Liu C, Fang T, Ou T, Wu C, Li R, Lin Y, et al. Global DNA methylation , DNMT1 , and MBD2 in patients with rheumatoid arthritis. *Immunol Lett* 2011;135:96–99.

12. Glossop JR, Emes RD, Nixon NB, Packham JC, Fryer AA, Mattey DL, et al. Genome-wide profiling in treatment-naive early rheumatoid arthritis reveals DNA methylome changes in T- and B-lymphocytes. *Epigenomics* 2015.

13. Glossop JR, Haworth KE, Emes RD, Nixon NB, Packham JC, Dawes PT, et al. DNA methylation profiling of synovial fluid FLS in rheumatoid arthritis reveals changes common with tissue-derived FLS. 2015;7:539–551.

14. Whitaker JW, Shoemaker R, Boyle DL, Hillman J, Anderson D, Wang W, et al. An imprinted rheumatoid arthritis methylome signature reflects pathogenic phenotype. *Genome Med* 2013;5:40.

15. Bartok B, Firestein GS. Fibroblast-like synoviocytes: Key effector cells in rheumatoid

arthritis. *Immunol Rev* 2010;233:233–255.

16. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum* 1988;31:315–324.

17. Barton JL, Trupin L, Schillinger D, Gansky SA, Tonner C, Margaretten M, et al. Racial and ethnic disparities in disease activity and function among persons with rheumatoid arthritis from university-affiliated clinics. *Arthritis Care Res* 2011;63:1238–1246.

18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559–575.

19. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006;2:2074–2093.

20. Consortium TIH. The International HapMap Project. *Nature* 2003;426:789–796.

21. Davis S, Du P, Bilke S, Jr Triche T, Bootwalla M. methylumi: Handle Illumina methylation data. 2014.

22. Yousefi P, Huen K, Schall RA, Decker A, Elboudwarej E, Quach H, et al. Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. *Epigenetics* 2013;8:1141–1152.

23. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 2013;29:189–196.

24. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013;8:203–209.

25. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B* 1995;57:289 – 300.

26. Rahmani E, Zaitlen N, Baran Y, Eng C, Hu D, Galanter J, et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat Methods* 2016;13:443–445.

27. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol* 2013;31:142–7.

28. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 2001;60:155–66.

29. Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. *Nat Methods* 2014;11:309–11.

30. Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet*

2012;44:1336–40.

31. Yarwood A, Han B, Raychaudhuri S, Bowes J, Lunt M, Pappas D a, et al. A weighted genetic risk score using all known susceptibility variants to estimate rheumatoid arthritis risk. *Ann Rheum Dis* 2013:1–7.

32. Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum* 1987;30:1205–1213.

33. Holoshitz J. The rheumatoid arthritis HLA-DRB1 shared epitope. *Curr Opin Rheumatol* 2010;22:293–298.

34. Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee H-S, Jia X, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet* 2012;44:291–296.

35. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;12:77.

36. Helm-Van Mil a. HM Van Der, Verpoort KN, Breedveld FC, Huizinga TWJ, Toes REM, Vries RRP De. The HLA-DRB1 shared epitope alleles are primarily a risk factor for anti-cyclic citrullinated peptide antibodies and are not an independent risk factor for development of rheumatoid arthritis. *Arthritis Rheum* 2006;54:1117–1121.

37. Stefansson OA, Moran S, Gomez A, Sayols S, Arribas-Jorba C, Sandoval J, et al. A DNA methylation-based definition of biologically distinct breast cancer subtypes. *Mol Oncol* 2015;9:555–568.

38. Tsai PC, Bell JT. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *Int J Epidemiol* 2015;44:1429–1441.

39. Georgiadis P, Hebels DG, Valavanis I, Liampa I, Bergdahl IA, Johansson A, et al. Omics for prediction of environmental health effects: Blood leukocyte-based cross-omic profiling reliably predicts diseases associated with tobacco smoking. *Sci Rep* 2016;6:20544.

40. Parkes M, Cortes A, Heel D a van, Brown M a. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet* 2013;14:661–73.

41. Cope AP, Schulze-Koops H, Aringer M. The central role of T cells in rheumatoid arthritis. *Clin Exp Rheumatol* 2007;25:S4–S11.

42. Fujii H, Shao L, Colmegna I, Goronzy JJ, Weyand CM. Telomerase insufficiency in rheumatoid arthritis. *Proc Natl Acad Sci U S A* 2009;106:4360–4365.

43. Goronzy JJ, Weyand CM. Rheumatoid arthritis. *Immunol Rev* 2005;204:55–73.

44. Shao L, Fujii H, Colmegna I, Oishi H, Goronzy JJ, Weyand CM. Deficiency of the DNA repair enzyme ATM in rheumatoid arthritis. *J Exp Med* 2009;206:1435–1449.

45. Yang Z, Fujii H, Mohan S V, Goronzy JJ, Weyand CM. Phosphofructokinase deficiency impairs ATP generation, autophagy, and redox balance in rheumatoid arthritis T cells. *J Exp Med* 2013;210:2119–34.

46. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén SE, Greco D, et al. Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. *PLoS One* 2012;7.

47. Plant D, Wilson AG, Barton A. Genetic and epigenetic predictors of responsiveness to treatment in RA. *Nat Rev Rheumatol* 2014;10:329–37.

48. Kim Y, Logan JW, Mason JB, Roubenoff R. DNA hypomethylation in inflammatory arthritis : Reversal with methotrexate. 1996:165–172.

49. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. *Genomics* 2011;98:288–295.

50. Illumina. Infinium HumanMethylation450 BeadChip Kit. 2015.

Tables and Figures

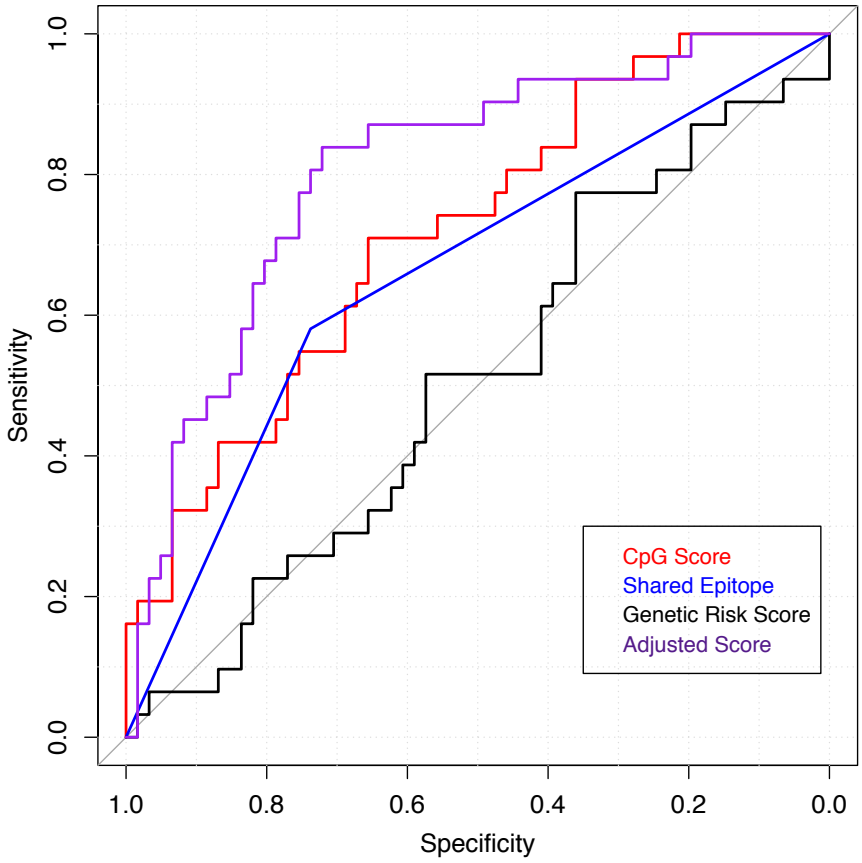


Figure 1. ROC curves of hypermethylation score, *HLA-DRB1* shared epitope, and genetic risk score as predictors of RA case status. The hypermethylation score is the sum of the beta values across the 1,056 significant CD4+ naïve T cell sites and is a measure of hypermethylation. Shared epitope is a binary variable taking on the value of 1 if a person has 1 or 2 copies of the shared epitope. Genetic risk score is a weighted score of 43 SNPs, previously validated (30,31). The adjusted hypermethylation score represents the sum of the 79 CpGs that were significant ($p<0.05$) after adjusting for age, smoking, batch and ReFACTor PC1.

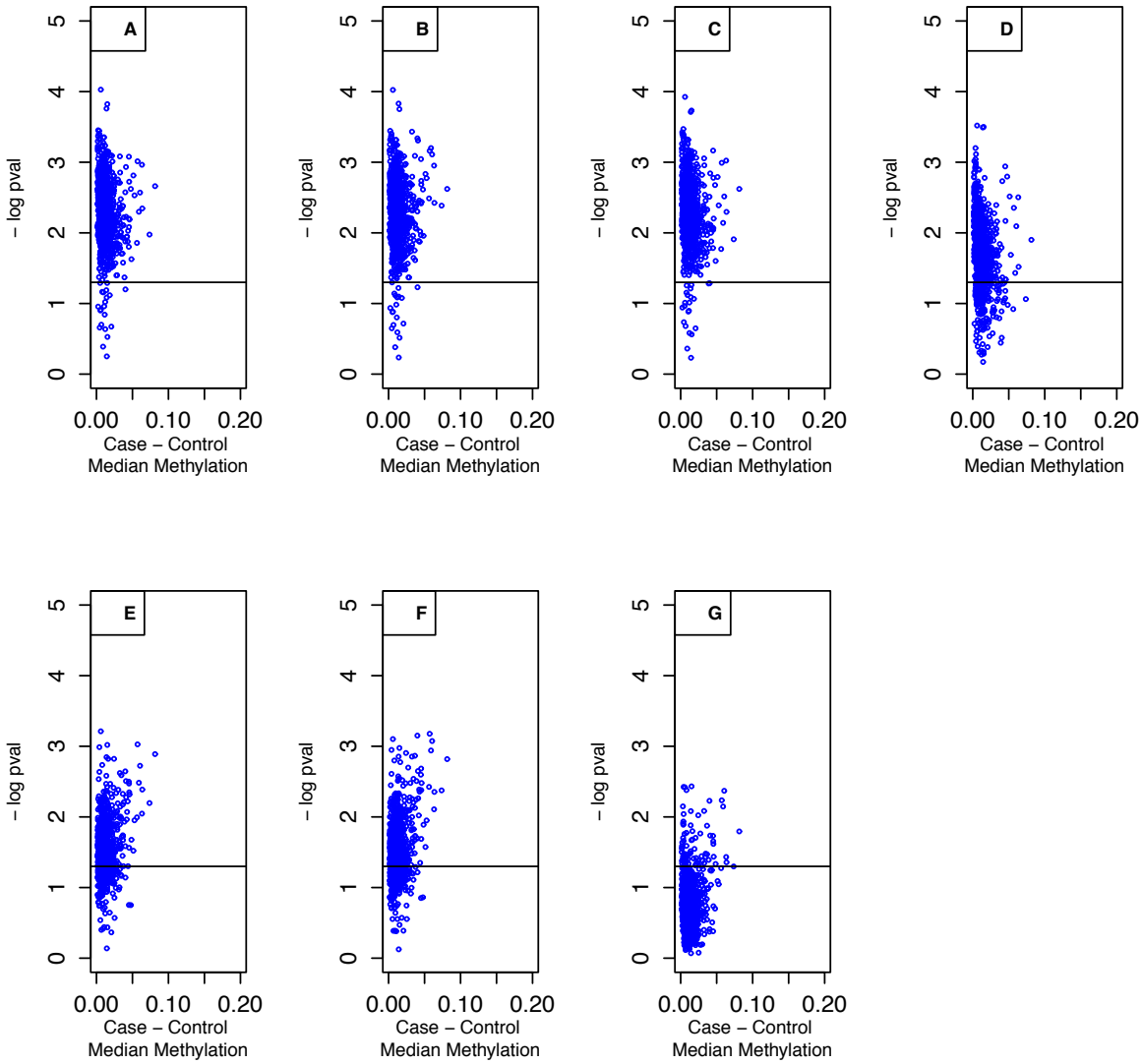


Figure 2A-G. Plots of one-sided p-values vs. median methylation difference in cases and controls for FLS CpGs in logistic regression models after adjusting for covariates. CpGs in the models are those significant at q<0.05 in Wilcoxon rank sum tests. Models are adjusted for (A) no covariates, (B) age only, (C) smoking only, (D) batch only, (E) ReFACTor PC1 only, (F) age, smoking and batch together, and (G) age, smoking, batch, and ReFACTor PC1 together, respectively.

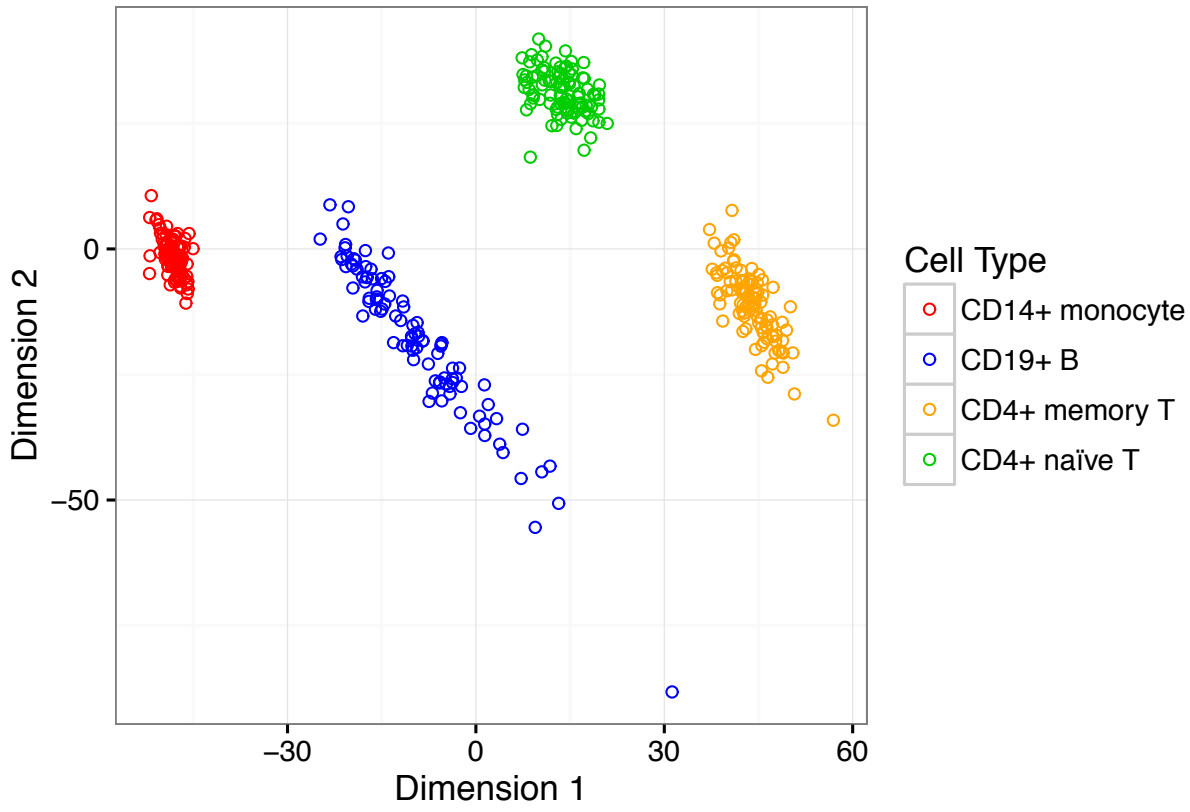


Figure 3. Multidimensional scaling (MDS) plot of DNA methylation profiles. This MDS plot of 371 samples (four immune cell types each for RA cases and controls) shows that samples cluster according to cell type, as expected.

Table 1. Study Participant Characteristics at Time of Blood Draw			
Characteristic	Cases (mean +/- SD) or count (%) n=63	Controls (mean +/- SD) or count (%) n=31	P-value (Wilcoxon or Chi Square)
Seropositive (RF[A] or CCP positive[B])	57 (90%)	----------	----------
Age	56.4 +/- 14.8	57.5 +/- 16.5	0.77
Smoking (Ever, Never)	33 (52%), 30 (48%)	13 (42%), 18 (58%)	0.34
Smoking (Current, Not Current)	4 (6%), 59 (94%)	1 (3%), 30 (97%)	0.53
Disease duration, years	14.0 +/- 10.5	----------	----------
Erosive disease (present, absent, missing)	39 (62%), 22 (35%), 2 (3%)	----------	----------
Disease activity: CDAI[C]	10.1 +/- 9.2 NA[D]=4	----------	----------

Table 1. Study Participant Characteristics at Time of Blood Draw. This table summarizes study participant characteristics at the time of blood draw.
[A]Rheumatoid Factor
[B]Anti-cyclic Citrullinated Peptide
[C]Clinical Disease Activity Index
[D]Not Available

Table 2. Candidate FLS CpG Results					
Cell Type	Raw p<0.05	Absolute median diff > 10%[A]	Absolute median diff between 1% and 10%[A]	FDR q<0.05	FDR q<0.05 and median diff > 1%
CD14 Hypomethylated	263	4	175	0	0
CD14 Hypermethylated	100	0	61	0	0
CD19 Hypomethylated	96	1	59	0	0
CD19 Hypermethylated	1408	1	732	0	0
CD4 Memory Hypomethylated	262	0	204	0	0
CD4 Memory Hypermethylated	66	1	36	0	0
CD4 Naïve Hypomethylated	160	1	62	0	0
CD4 Naïve Hypermethylated	2,569	0	1,105	1,056	517

Table 2. Candidate FLS CpG Results. Wilcoxon rank sum tests were carried out for each FLS candidate CpG in each of the four cell types with one-sided p-values, according to whether the CpG was hypermethylated or hypomethylated in the original study.
[A]Absolute median difference numbers are among the CpGs with unadjusted p<0.05

Table 3. ROC Areas Under the Curve	
Model	AUC (95% CI)
Hypermethylation Score (1,056 sites)	72% (61%-83%)
Shared Epitope	66% (56%-76%)
Genetic Risk Score	51% (38%-63%)
Hypermethylation Score (830 sites; Age, Smoking, Batch Adjusted Regression)	72% (61%-83%)
Hypermethylation Score (79 sites; Age, Smoking, Batch, ReFACTor PC1 Adjusted Regression)	81% (71%-90%)

Table 3. ROC Areas Under the Curve. ROC analysis was carried out for a hypermethylation score based on the 1,056 CpG sites significant at $q<0.05$ from the Wilcoxon rank sum tests. This score was compared to shared epitope status (positive/negative) and a genetic risk score. Two other hypermethylation scores were constructed, based on the 1,056 CpGs that remained significant ($p<0.05$) in logistic regression models after adjusting for various covariates.

Supplementary Materials

Supplementary Text 1: Results of Overnight Cell Storage

PBMCs were isolated and stained for FACS on the same day as blood collection, then either sorted the same day or stored overnight at 4°C before sorting. The impact of storing cells overnight was assessed for five outcomes: cell count, purity, quantitative DNA yields, DNA quality and stability of DNA methylation profiles. High purity (total >90%; gated >96%) was observed for all sorted cell populations, regardless of whether FACS occurred on the same day or the day following blood collection. The number of cells collected ranged from 1.4-4.6 million cells for samples sorted on the same day and 1.7-5.7 million cells for samples sorted the next day, and the numbers were similar for all cell types. High quality DNA (260/280 ratio: 1.87-1.95) was obtained from all cell types (3.2-39.0 ug). Paired DNA samples from all four cell types were collected at both time points. Overall, correlation between profiles derived from all paired samples for the four cell types was very high (r^2>0.997).

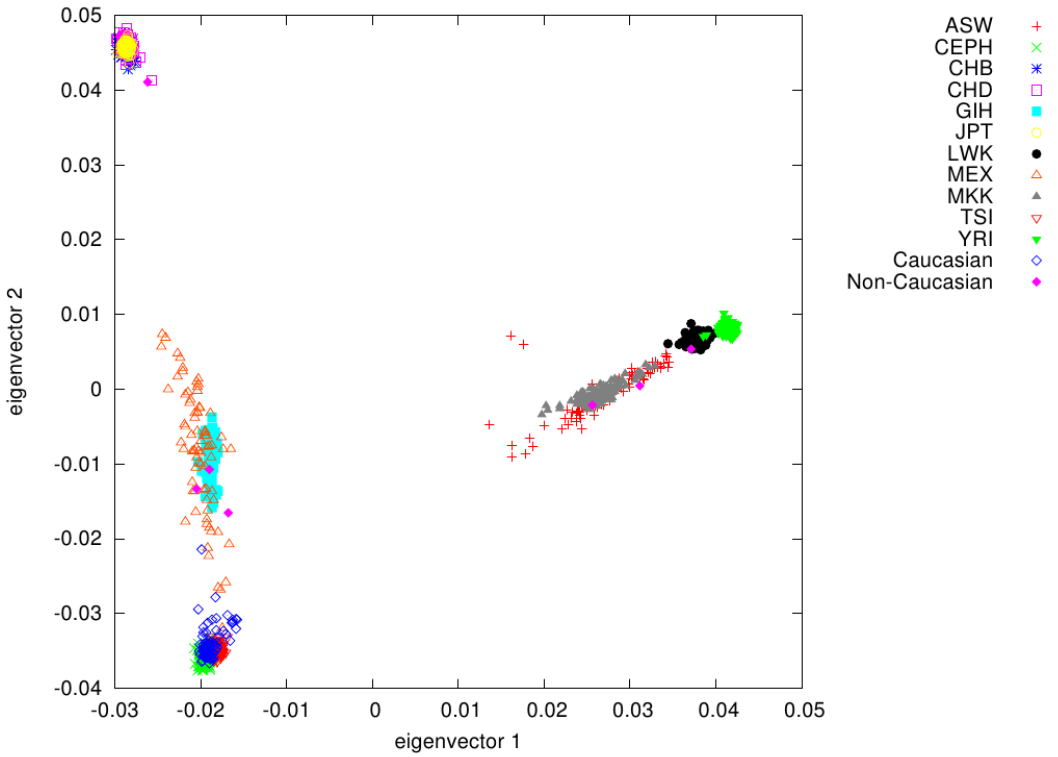


Figure S1. Eigenvector 1 and 2 from EIGENSTRAT in all samples. Eigenvector 1 vs Eigenvector 2 from EIGENSTRAT, showing where the RA European-ancestry and non-European-ancestry samples cluster relative to HapMap populations, prior to removal of non-European-ancestry samples[44].

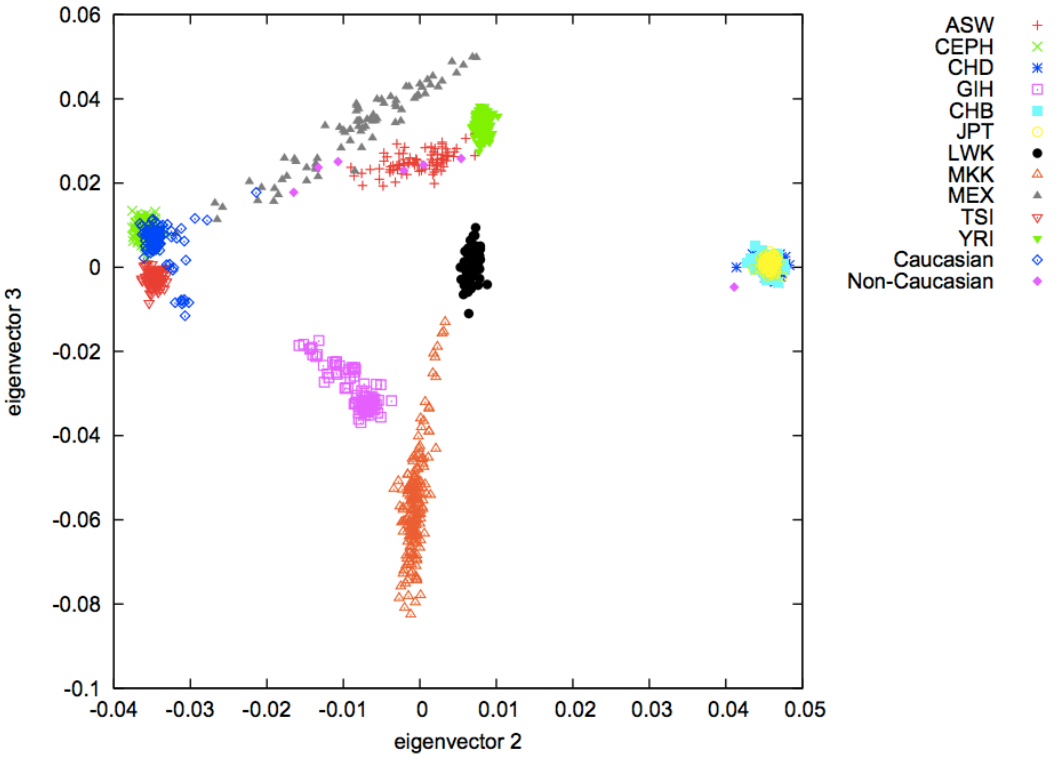


Figure S2. Eigenvector 2 and 3 from EIGENSTRAT in all samples. Eigenvector 2 vs Eigenvector 3 from EIGENSTRAT, showing where the RA European-ancestry and non-European-ancestry samples cluster relative to HapMap populations, prior to removal of non-European-ancestry samples[44].

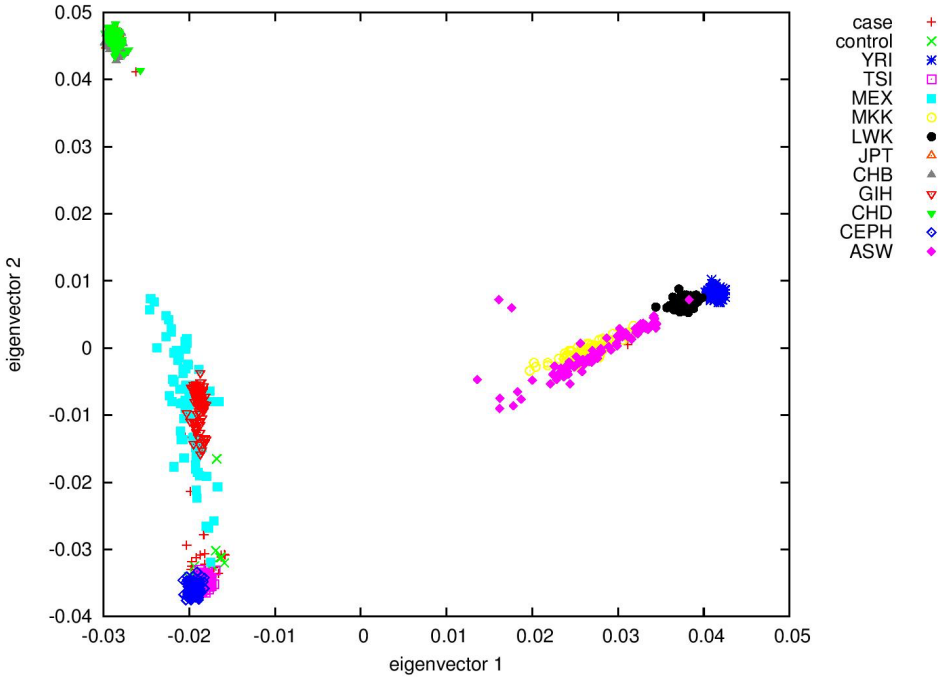


Figure S3. Eigenvector 1 and 2 from EIGENSTRAT in European ancestry samples. Eigenvector 1 vs Eigenvector 2 from EIGENSTRAT, showing where the RA European-ancestry cases and controls samples cluster relative to HapMap populations[44].

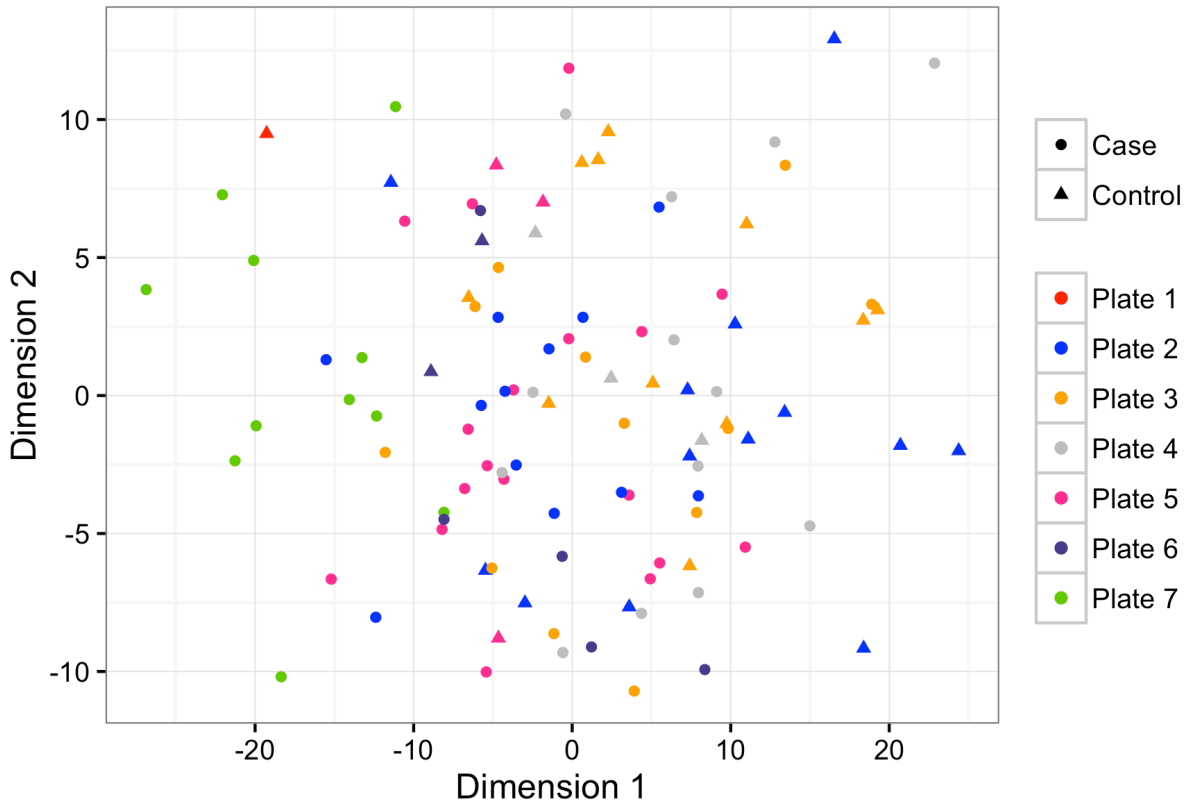


Figure S4. MDS of CD4+ naïve T cells before normalization. Multidimensional scaling (MDS) plot for CD4+ naïve T cells showing samples colored by batch, before background subtraction and normalization. C1 is MDS component 1 and C2 is MDS component 2.

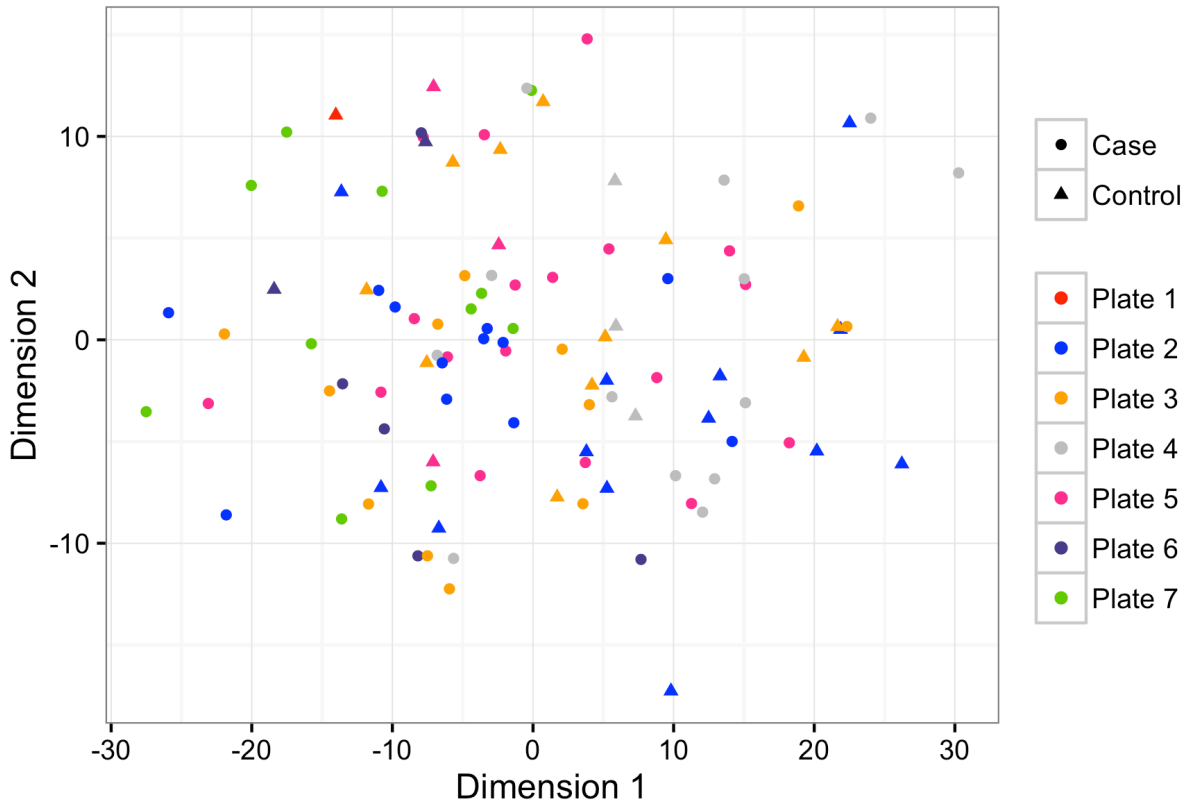


Figure S5. MDS of CD4+ naïve T cells after normalization. Multidimensional scaling (MDS) plot for CD4+ naïve T cells showing samples colored by batch, after background subtraction and normalization. C1 is MDS component 1 and C2 is MDS component 2.

Table S1. FLS Candidate Sites Replicated in CD4 Naïve Cells in Peripheral Blood of RA Cases. This table lists the FLS candidate sites that were significant in our study ($q<0.05$) in CD4+ naïve T cells. (Separate file chapter2_supplementary_table_S1.xlsx.)

Table S2. Logistic Regression Results from Wilcoxon Rank Sum Test FLS Sites	
Model adjusted for	Number CpGs p<0.05
No Covariates	1,035
Age	1,036
Smoking	1,034
Batch	841
ReFACTor PC1	837
Age, Smoking, Batch	830
Age, Smoking, Batch, ReFACTor PC1	79

Table S2. Logistic Regression Results from Wilcoxon Rank Sum Test FLS Sites. Logistic regression models were carried out on RA status adjusting for each of the 1,056 CpGs that were significant at q<0.05 in the Wilcoxon rank sum tests, and adjusting for various covariates. For each CpG, individual models adjusting for these covariates were performed.

Table S3. Candidate CpGs from Genes Previously Associated with RA. This table lists the candidate CpGs that are within genes previously associated with RA. (Separate file chapter2_supplementary_table_S3.xlsx.)

Table S4. Candidate GWAS CpG Results					
Cell Type	Raw p<0.05	FDR q<0.05 (direction)	Absolute median diff > 10%[a]	Absolute median diff between 1% and 10%[a]	FDR q<0.05 and median diff > 1%
CD14	77	0	1	165	0
CD19	225	0	1	440	0
CD4 Memory	65	1 (hyper)	0	214	0
CD4 Naïve	480	6 (hypo) 12 (hyper)	0	357	2

Table S4. Candidate GWAS CpG Results. A two-tailed Wilcoxon rank sum test was carried for each of the 1,676 CpGs from genes previously associated with RA (4) to compare the median methylation value between RA cases and controls, for each of the four cell types in peripheral blood. Results were corrected for multiple testing. Median difference is the median β methylation value in cases for a CpG minus the median β methylation value in controls for that CpG.

[A]Absolute median difference numbers are among the CpGs with unadjusted p<0.05

Table S5. Logistic Regression Results Using 18 Significant CpGs from Wilcoxon Rank Sum Test, GWA Sites	
Model adjusted for	Number CpGs p<0.05
No Covariates	18
Age	18
Smoking	18
Batch	16
ReFACTor PC1	18
Age, Smoking, Batch	16
Age, Smoking, Batch, ReFACTor PC1	11

Table S5. Logistic Regression Results Using 18 Significant CpGs from Wilcoxon Rank Sum Test, GWA Sites. Logistic regression models were carried out on RA status adjusting for each CpG that was significant at q<0.05 in the Wilcoxon rank sum tests, and adjusting for various covariates. For each CpG, individual models adjusting for these covariates were performed.

Table S6. Genome-wide Results				
Cell Type	Raw p<0.05	FDR q<0.05 (% of total)	Absolute median diff > 10%[A]	Absolute median diff between 1% and 10%[A]
CD14	17487	0	45	4342
CD19	57493	0	43	25806
CD4 Memory	10793	0	31	3708
CD4 Naïve	10793	0	31	3708

Table S6. Genome-wide Results. A 2-tailed Wilcoxon rank sum test was performed for each of the 428,232 CpGs on the 450k BeadChip following quality control to compare the median methylation value between RA cases and controls, for each of the four cell types in peripheral blood. Multiple testing was accounted for by controlling the false discovery rate (FDR). Median difference is the median β methylation value in RA cases for a CpG minus the median β methylation value in controls for that CpG.
[A]Absolute median difference numbers are among the CpGs with unadjusted p-value of <0.05

Table S7. Mean Methylation in RA Cases and Controls.			
Cell Type	Mean of RA Cases	Mean of Controls	P-value
CD14	0.501	0.500	0.81
CD19	0.513	0.509	0.32
CD4 Memory	0.502	0.503	0.92
CD4 Naïve	0.528	0.523	0.06

Table S7. Mean Methylation in RA Cases and Controls. A global mean methylation level in each sample was determined by finding the mean β methylation value for all sites that passed QC filtering for that sample, across all CpGs for each RA case and control. For each cell type a t-test was calculated to determine whether global methylation differed between cases and controls.

Chapter 3 - Increased DNA methylation of *SLFN12* in CD4+ and CD8+ T cells from multiple sclerosis patients

Abstract

DNA methylation is an epigenetic mark that is influenced by environmental factors and is associated with changes to gene expression and phenotypes. It may link environmental exposures to disease etiology or indicate important gene pathways involved in disease pathogenesis. We identified genomic regions that are differentially methylated in T cells of patients with relapsing remitting multiple sclerosis (MS) compared to healthy controls. DNA methylation was assessed at 450,000 genomic sites in $CD4^+$ and $CD8^+$ T cells purified from peripheral blood of 94 women with MS and 94 healthy women, and differentially methylated regions were identified using *bumphunter*. Differential DNA methylation was observed near four loci: *MOG*/*ZFP57*, *NINJ2/LOC100049716, HLA-DRB1*, and *SLFN12*. Increased methylation of the first exon of the *SLFN12* gene was observed in both T cell subtypes and remained present after restricting analyses to samples from patients who had never been on treatment or had been off treatment for more than 2.5 years. Genes near the regions of differential methylation in T cells were assessed for differential expression in whole blood samples from a separate population of 1,329 women with MS and 97 healthy women. Gene expression of *HLA-DRB1, NINJ2*, and *SLFN12* was observed to be decreased in whole blood in MS patients compared to controls. We conclude that T cells from MS patients display regions of differential DNA methylation compared to controls, and corresponding gene expression differences are observed in whole blood. Two of the genes that showed both methylation and expression differences, *NINJ2* and *SLFN12*, have not previously been implicated in MS. *SLFN12* is a particularly compelling target of further research, as this gene is known to be down-regulated during T cell activation and up-regulated by type I interferons (IFNs), which are used to treat MS.

Introduction

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system, with onset during early adulthood, leading to demyelination and axonal degeneration that often progresses to physical and cognitive disability. The cause of MS is unknown, however, genetic and environmental factors, and interactions between them, are known to contribute to disease risk.[1–3] Variation in human leukocyte antigen (HLA) genes represent the strongest genetic susceptibility factor for MS, with the strongest signal in *HLA-DRB1*. In recent years, genome-wide association studies (GWAS) and custom chip-based studies have identified 200 MS-associated non-HLA loci.[4–6] Each of these genetic associations exerts only a modest effect size, and no genetic variant by itself is sufficient to cause MS, making the genetic contribution to MS etiology highly complex. The local linkage disequilibrium (LD) structure of most MS-associated loci makes the identification of true causal variants difficult. However, when inferring the most likely affected genes, a strong overrepresentation of immunologically relevant genes is observed, in particular for genes known to regulate T cell mediated immunity.[4,6]

MS heritability is not yet fully explained through the associated genetic variants, indicating that additional factors, such as epigenetic mechanisms, contribute to MS etiology. The term epigenetics describes heritable changes in gene regulation that do not alter the DNA sequence. DNA methylation, a widely studied epigenetic mechanism, is the addition of a methyl group to the fifth carbon position of cytosine at CpG dinucleotides. DNA methylation in gene promoter regions typically prevents transcription factors from binding and thereby silences gene expression, although other regulatory effects of DNA methylation are known.[7] While DNA methylation patterns can be inherited, they are also affected by environmental exposures such as tobacco smoke, diet, exercise, stress, and medications. Thus, DNA methylation may link environmental exposures and genetic variations to MS disease risk. DNA methylation associations have been shown in cancers,[8] and more recently, immune-mediated and neurodegenerative diseases,[9–11] including MS.[12–20]

Here, we investigate DNA methylation in $CD4^+$ and $CD8^+$ T cells purified from blood in Norwegian and Australian MS patients compared to healthy controls. Methylation differences in these cell types between MS patients and controls have been previously studied in both smaller cohorts.[13,15,16,20] More samples have since been added, and the Australian and Norwegian datasets have been combined to maximize statistical power. The current analysis represents the largest study to date on the role of DNA methylation of immune cells in MS. Epigenome-wide association analysis was performed to identify differentially methylated positions (DMPs) and differentially methylated regions (DMRs). Gene expression changes in whole blood corresponding to DMRs was used to validate our methylation findings in an independent dataset of MS cases and healthy controls.

Results

Participant characteristics and proportion of cell type samples available for analyses are summarized in Table 1. $CD4^+$ and $CD8^+$ T cells were analyzed separately, and different subsets of cases were considered to evaluate potential bias based on treatment. In total, five sub-analyses were conducted: a) $CD4^+$ T cells of all cases regardless of treatment vs. all controls; b) $CD8^+$ T cells of all cases regardless of treatment vs. all controls; c) $CD4^+$ T cells of cases not on treatment at the time of inclusion vs. all controls; d) $CD4^+$ T cells of treatment-naïve cases vs. all controls; and e) $CD8^+$ T cells of treatment-naïve cases vs. all controls (Table 2). There were insufficient $CD8^+$ T cell samples to analyze cases off-treatment at time of inclusion. About 65,000 CpGs were removed from each dataset in quality control steps. The genomic inflation factor was close to one for all analysis strata, indicating that results were not considerably confounded after including surrogate variables (SVs) in the regression models. We compared estimated SVs to the measured variables of participant age and batch and found that each of these features was well captured in the largest SVs (S1 and S2 Figs).

DMP analysis confirms hypermethylation in $CD8^+$ T cells for MS patients

No individual DMPs were significantly associated with MS after adjusting for multiple hypothesis testing. However, when we focused on probes that showed a nominally significant p-value in the DMP analysis of all samples, we confirmed our previous findings[16] that $CD8^+$ T cells of MS patients display a higher degree of DNA methylation as compared to healthy controls (Fig 1). This trend becomes increasingly apparent as p-values become increasingly stringent, ranging from 52% of sites hypermethylated at $p<0.05$ to 69% hypermethylated at $p<0.0001$. In $CD4^+$ T cells no trend towards DNA hypermethylation was observed for any p-value cutoff.

DMRs in MS patients compared to controls

As groups of CpG sites located near one another can be methylated or demethylated together, and identifying these regions of differential methylation is statistically more powerful than identifying single DMPs, we next sought to identify DMRs.[21] Results are summarized in Table 3. The exact same DMRs were identified for $CD4^+$ T cells of cases not on treatment at the time of inclusion and $CD4^+$ T cells of treatment-naïve cases (datasets c and d listed above), so only results for the latter are included here. Additionally, because microarray probes used to assess DNA methylation may be sensitive to SNPs in the probe sequences, we evaluated whether methylation at individual CpGs within DMRs corresponded to differences in genotypes. Out of 34 CpGs in DMRs with SNPs in the probe sequences that were also present in the imputed Norwegian genetic data, 4 CpGs in the *MOG*/*ZFP57* DMR were found to be differentially methylated by genotype. Dropping the 4 CpG sites resulted in a slightly higher family-wise error rate (FWER) for this DMR, but the result remained significant.

Hypermethylation of SLFN12 is associated with MS

A consistent DMR signal was observed on chromosome 17 in $CD4^+$ and $CD8^+$ T cells (Table 3). A long DMR (between 18 and 22 differentially methylated CpGs, depending on the dataset analyzed) covering the first exon of *SLFN12* showed hypermethylation in MS patients compared to healthy controls in both $CD4^+$ and $CD8^+$ T cells (Fig 2). Hypermethylation was seen in all strata, regardless of treatment status of cases. A much smaller DMR consisting of a single CpG site hypomethylated only in the $CD4^+$ T cells of treatment-naïve cases compared to controls was identified 3kb downstream of *SLFN12* (Table 3). We note that a DMR can consist of a single CpG site due to the width of the *Bumphunter* smoothing function.

Hypomethylation in the MHC region

Evidence for a DMR was observed in a regulatory region just outside the HLA Class I region on chromosome 6 in $CD4^+$ T cells (Table 3). Specifically, this DMR is located in a regulatory region 8kb downstream of *MOG*, encoding myelin oligodendrocyte glycoprotein, which is expressed in myelin sheaths,[22] and 3kb upstream of the zinc finger protein gene *ZFP57*, encoding a protein that likely acts a as a transcriptional

repressor (RefSeq, Sep 2009). In addition, we confirmed evidence of hypomethylation in the *HLA-DRB1* gene in MS $CD4^+$ T cells compared to healthy controls, as previously reported,[20] as well as in $CD8^+$ T cells (Table 3). However, the *HLA-DRB1* result was not observed when analyses were restricted to treatment-naïve cases.

Hypermethylation of the NINJ2/LOC100049716 locus

A DMR consisting of 3 CpGs in the first intron of the *NINJ2* gene demonstrated hypermethylation in $CD4^+$ T cells from treatment-naïve MS patients when compared to healthy controls. This region is overlapped by the first exon of an uncharacterized long non-coding RNA (*LOC100049716*).

DMRs correspond to differential expression of genes in whole blood

To assess whether the observed DNA methylation was associated with gene expression of nearby genes, we nominated six candidates in close proximity to the DMRs identified in the current study: *SLFN12*, *NINJ2*, *MOG*, *HLA-DRB1, ZFP57,* and *LOC100049716*. Using whole blood samples from a large collection of MS patients and healthy controls, differential gene expression was assessed for four of the genes. *MOG* and *ZFP57* expression levels were below the microarray background threshold (average log2 expression <4) and therefore not considered in our analyses. Lower gene expression was observed for *SLFN12, HLA-DRB1* and *NINJ2* in MS patients compared to healthy controls, and there was no difference in *LOC100049716* (Fig 3). Results from genome-wide analysis (*limma*) and individual linear regression fits are listed in Table 4.

Discussion

Our findings show that $CD4^+$ and $CD8^+$ T cells isolated from MS patients have regions of markedly increased or decreased methylation compared to cells isolated from healthy controls. These regions may influence disease etiology and shed light on risk factors for MS.[3] Compellingly, a DMR flanking the first exon of *SLFN12* occurred in all patient subsets for both $CD4^+$ and $CD8^+$ T cells. *SLFN12* encodes a member of the Schlafen protein family, which is a family of proteins encoded by a cluster of five genes on chromosome 17. Type I IFNs induce the expression of Schlafen genes.[23] *SLFN12* has been shown to be downregulated during T-cell activation in primary human cells.[24] From clinical observations and genetic studies,[4,6,25–27] there is convincing evidence that MS pathology is driven by T-cells, and IFN beta type I is an approved therapy for MS, making *SLFN12* a biologically plausible gene of interest for MS. Experimental evidence shows that *Slfn8* knockout mice have lower expression of pro-inflammatory cytokines and are resistant to induced experimental autoimmune encephalomyelitis (EAE), the mouse model of MS.[28] Though *Slfn8* is a different member of the Schlafen family, its clear role in EAE makes *SLFN12* an appealing target for further research in MS.

Hypermethylation of the first exon of *SLFN12* suggests repression of this gene in samples from MS patients, which is corroborated by decreased expression in whole blood of MS

patients compared to controls (Table 4) in an independent cohort. It is not known whether this decrease in gene expression is caused by the observed hypermethylation or is a result of increased T-cell activity. A study of the transcriptional co-regulator gene *Mastermind-Like 1 (MAML1)* showed that its overexpression in embryonic kidney cells induced widespread methylation changes, including hypermethylation of *SLFN12* and corresponding downregulation of the gene, suggesting that a change in methylation alone could be responsible for decreased expression in T cells.[29] Conversely, a study in allergic rhinitis sufferers also found increased methylation and decreased expression of *SLFN12* in lymphocyte-enriched blood after participants were exposed to allergens, suggesting that T-cell activation could be the primary instigating factor.[30] Additionally, the area of the *SLFN12* DMR is enriched in the H3K27Ac histone mark, which is typically associated with regulatory elements, and it contains over 50 transcription factor binding motifs (Fig 4); increased methylation would therefore be expected to result in decreased expression.[31,32]

In line with earlier findings by Graves *et al*.[13] and Maltby *et al*.,[20] we confirmed that *HLA-DRB1* is hypomethylated in the $CD4^+$ T cells of MS patients and observed the DMR for the first time in $CD8^+$ T cells. *HLA-DRB1* is highly polymorphic, and is the strongest genetic risk for MS, raising the question of whether the observed differential methylation could be attributable to genetic variation in probe sequences used on the Illumina array. However, no CpGs with SNPs in probe sequences with differential methylation by genotype in this DMR were found. Interestingly, the *HLA-DRB1* DMR was identified using a different method from that described by Maltby *et al*. When we investigated the gene expression of *HLA-DRB1*, we observed that while hypomethylation was present in MS patients compared to controls, this gene has decreased expression in MS patients. This finding could be due to the location of the DMR in the gene body rather than a promoter,[33] or due to the fact that gene expression was investigated in whole blood rather than isolated T cells. Of note, the DMR was only detected when including all MS patients in the analysis regardless of treatment and not when restricting to off-treatment patients. This finding could be explained by lower statistical power in the off-treatment subgroups, or the DMR could result from medications used to treat MS.

Finally, hypermethylation of a region near *NINJ2* was observed for $CD4^+$ T cells, which corresponded with lower expression of this gene in whole blood. This gene has previously been reported to show aberrant methylation in borderline personality disorder.[34] The DMR was only evident when off-treatment cases were included and was not detected when on-treatment cases were added, suggesting that treatment could be altering methylation in this region. This finding needs to be validated in an independent dataset.

When we compared the DMRs against the recently published 200 MS-associated SNPs[6] we did not observe any overlap outside of *HLA-DRB1*. It is possible that DNA methylation represents an independent functional mechanism of MS etiology. Though not in the list of 200 definite MS-associated SNPs, variants in *NINJ2* have been identified as “suggestive” MS-associated SNPs.[6] The increased methylation in this gene in $CD4^+$ T

cells from MS cases could be due to *NINJ2* variants, to an environmental factor, or to both, possibly acting in concert within an individual.

Some key strengths of this study were that searching for regions of differential methylation instead of isolated CpG sites provided greater statistical power, and that the relationships between our top DMRs and potential impact on mRNA levels were investigated a large independent dataset. In addition, the use of careful quality control procedures ensured that results were not due to technical artifacts. The use of SVA to infer covariates in the data allowed us to adjust for both measured and unmeasured confounders, and we confirmed that known variables such as measurement batch and BeadChip type were captured by SVs. Separate analysis of treatment-naïve cases allowed us to confirm that results were not due solely to use of medications. A limitation of this study was that DNA methylation was measured in cases after they developed MS, therefore temporality between methylation changes and disease onset could not be established. Also, environmental exposures were not evaluated. Methylation was assessed in T cells while expression was assessed in whole blood, which may not completely capture the relationship between DMRs and expression in T cells. Of the ~28 million CpG sites in the human genome, methylation was assessed only for sites on the 450k BeadChip.[35] Larger sample sizes will be needed to investigate differences in findings between $CD4^+$ and $CD8^+$ T cells. Finally, because this study was restricted to white females, findings may be sex specific and not generalizable to other populations.

Known environmental MS risk factors may exert effects on MS risk via changes in DNA methylation. For example, smoking increases MS risk, especially among carriers of HLA risk alleles or carriers of variants in *NAT1*.[3] Furthermore, smoking is associated with demethylation of the aryl hydrocarbon receptor repressor (*AHRR*) gene,[36] and the effect of smoking on demethylation of *AHRR* in blood is more pronounced in MS cases than healthy controls.[37] Larger effects from smoking on methylation throughout the genome have been observed in MS cases carrying *HLA-DRB1*15:01* and lacking the *HLA-A*02* protective variant, and demethylation of a single CpG site near *SLFN12L* in former smokers relative to never smokers was detected in these cases.[37] Smoking is also associated with a more severe disease course.[3] These studies suggest that methylation could potentially be a mechanism by which smoking is acting to alter risk of disease or severity of disease course.

Other well-established MS environmental risk factors are associated with altered DNA methylation patterns and could help explain our findings. Adiposity has been found to be the cause of genome-wide methylation changes,[38] and adolescent obesity is associated with a two-fold risk of MS. Low vitamin D and decreased sun exposure are also associated with MS, and vitamin D can alter methylation status of other genes.[39] Several studies have shown an association of Epstein-Barr virus (EBV) with MS, and EBV exploits the epigenetic machinery of infected host cells to regulate its life cycle, resulting in widespread methylation changes to the host cell.[40] However, EBV resides in epithelial and B cells, so if methylation changes are due to EBV, they are more likely to be seen in those cell types. The impact of these environmental factors on methylation patterns

specifically relevant to MS needs to be investigated further, in T cells and other immune cell types and tissues, as methylation changes may help explain the biological mechanisms through which environment affects disease risk, and consequently may identify new therapeutic targets that have not been revealed by genetic studies alone.

In conclusion, this is the largest genome-wide DNA methylation study of MS in $CD4^+$ and $CD8^+$ T cells to date. We show evidence that DNA methylation of $CD4^+$ and $CD8^+$ T cells plays a role in MS etiology. Consistent DMRs in *SLFN12* and *HLA-DRB1* were observed across two T cell sub-types, and differential gene expression was detected in whole blood for these gene candidates. Results indicate that DMRs may be detected in more accessible whole blood samples, paving the way for future large-scale studies of DNA methylation in MS. These findings would benefit from additional confirmation in larger independent case-control studies. Further research investigating the functional mechanisms underlying the association of these methylated regions to MS is warranted, particularly for *SLFN12*, which is not well-characterized.

Methods

Study populations

Norwegian patients with relapsing remitting MS diagnosed according to the McDonald criteria[41] (N = 46 females) were recruited from the Department of Neurology at Oslo University Hospital, Norway.[16] Controls (N = 46 females) were recruited through patients or from hospital employees and were frequency-matched to cases by age in 5-year increments. Almost all the Norwegian patients were treatment-naïve at the time of inclusion, except two patients were on IFN beta treatment, and three patients had previously received medications (none antibody-based, with a washout time of at least 2.5 years prior to inclusion).

Australian relapsing remitting MS patients were recruited from the John Hunter Hospital MS Clinic in New South Wales, Australia, and controls were recruited from the Australian Red Cross Blood Bank.[15,20] The Australian patients were a mix of patients who were treatment-naïve, off treatment for >3 months, or on treatment at the time of inclusion.

The Norwegian Regional Committee for Medical and Health Research Ethics and the Australian Hunter New England Health Research Ethics (05/04/13.09) and University of Newcastle Ethics (H-505-0607) committees approved this study. Methods were carried out in accordance with institutional guidelines on human subject experiments. Written and informed consent was obtained from all subjects.

Purification of $CD4^+$ and $CD8^+$ T cells

For the Norwegian samples, $CD4^+$ and $CD8^+$ T cells were isolated from freshly collected peripheral blood mononuclear cells (PBMCs) using immunomagnetic cell separation selection kits (EasySep Human $CD4^+$ T cell Isolation Kit (negative selection) and

EasySep Human $CD8^+$ Selection Kit (positive selection), StemCell Technologies, Canada) according to the manufacturer instructions. Purified cells were stained with FITC-conjugated mouse anti-human CD4 (clone RFT4, catalog #9522-02, Southern Biotech, USA) or FITC-conjugated mouse anti-human CD8 (clone HIT8a, catalog #555634, BD Biosciences, USA) and FITC-conjugated mouse IgG1 isotype control (clone 15H6, catalog #0102-02, Southern Biotech, USA) antibodies, and purity exceeding 95% was confirmed by flow cytometry (Attune Acoustic Focusing Cytometer, Applied Biosystems, USA).

For the Australian samples, $CD4^+$ and $CD8^+$ T cells were isolated from PBMCs using the same methods as the Norwegian samples. The purity of the cells was assessed by flow cytometry using a FITC conjugated anti-human CD4 antibody (clone OTK4, catalog #60016FI, StemCell Technologies, Canada) or an anti-human CD8 antibody (clone RPA-T8, catalog #60022FI.1, StemCell Technologies, Canada) on a BD FACSCanto II flow cytometer, then analyzed using FACSDiva software (BD Biosciences, USA) at the Analytical Biomolecular Research Facility of the University of Newcastle. All samples met a minimum purity threshold of >90%.

DNA extraction

DNA from purified $CD4^+$ and $CD8^+$ T cell samples was extracted using QIAamp DNA Mini Kit (Qiagen, Germany) and bisulfite converted with the EZ DNA Methylation Kit (Zymo Research, USA). Methylation was assayed using Illumina BeadChips according to the manufacturer instructions (Illumina, USA). Two thirds of the Norwegian cohort were assayed with MethylationEPIC (EPIC) BeadChips. The Australian cohort and the rest of the Norwegian cohort were assayed with HumanMethylation450 (450k) BeadChips. Norwegian samples were assayed in four batches; Australian samples were assayed in two batches.

Genotyping and imputation

The Norwegian samples were genotyped with the Human Omni Express BeadChip (Illumina). PLINKv1.09[42] was used to apply consecutive filters for per-SNP call rate (0.95) and per-sample call rate (0.95) prior to pre-phasing and imputation. MACH[43] was used for pre-phasing and genotypes were imputed against European samples from the 1000 Genomes data release 3 using Minimac3.[44] The Australian samples were not genotyped.

DNA methylation data processing and analysis

The *Minfi* R package was used for pre-preprocessing, normalization, and quality control (QC).[45] EPIC and 450k datasets were combined by extracting the probes present on both platforms. The Norwegian and Australian samples were combined, and then datasets were split by cell type. Processing and analyses of $CD4^+$ and $CD8^+$ T cells were performed separately. Separate analyses were performed for a) treatment-naïve patients;

b) those off treatment for >3 months at the time of inclusion in addition to treatment-naïve patients; and c) all patients. The *preprocessNoob* function was used for background subtraction and dye normalization, followed by quantile normalization with *preprocessQuantile*. Data points with detection p-value > 0.01 were replaced with "NA" values. CpG sites with more than 5% "NA" values across all samples were discarded. CpG sites with a common SNP (based on hg19, dbSNP build 144) at the CpG interrogation site or the single base extension were removed from analysis, as were sites with probes predicted to cross-hybridize to other genomic locations.[46] Predicted gender based on X and Y chromosome methylation matched each study participant's gender. None of the samples had more than 5% "NA" values for sites forwarded for analysis.

Methylation beta values, defined as the proportion the methylated signal makes up of the methylated plus unmethylated signal, were logit transformed into M-values for analysis to reduce heteroskedasticity. Surrogate variable analysis (SVA) was run on each separate dataset analysis stratum to find latent variables. Such variables may represent batch effects, cellular heterogeneity or other unknown confounders (e.g., varying levels of inflammation). Measured potential covariates, such as age and batch, were not included when estimating SVs to allow SVA to identify such covariates directly from the data, as described by Jaffe *et al.*[47] and Leek,[48] with the "be" option used to determine the number of SVs to calculate, since this option resulted in lower genomic inflation compared to the "leek" option.[48] To determine DMPs, an epigenome-wide association analysis was performed using the empirical Bayes method in *limma*, with the M-value for each CpG as the outcome variable and disease status and SVs as predictors.[49]

Differentially methylated regions (DMRs) were identified using *Bumphunter*,[50] using the same outcome and predictor variables as in the DMP analysis. CpG sites separated by at most 500 bp were used to define clusters, and then 1,000 bootstrap samples were used to generate a null distribution of regions. Candidate regions were nominated with *pickCutoff*, using the 99% quantile of the null-distribution as a threshold. An adjusted p-value cutoff of 0.2 for the family-wise error rate (FWER) produced by *Bumphunter* was used to assign DMRs in each analysis. A liberal p-value cutoff was used because controlling the FWER (the probability of making at least one type I error) is more conservative than controlling the false discovery rate (which controls the proportion of type I errors). Imputed genome-wide SNP data was used to identify probes containing polymorphic SNPs in their recognition sequence for the Norwegian data. For each DMR, CpG sites with probes containing SNPs were further assessed for differential methylation by genotype. Those CpG sites were then excluded and supplementary DMR analyses were performed for the entire dataset.

Whole blood gene expression data generation and pre-processing

Gene expression was evaluated in a separate population. Whole blood PAXgene tubes were collected from 1,329 female relapsing remitting MS patients at baseline of the phase 3 studies DEFINE and CONFIRM for demonstrating efficacy of delayed-release dimethyl fumarate for the treatment of RRMS, and from 97 female healthy volunteers.[51,52]

These subjects were predominantly of European ancestry (87%) and treatment-naïve (77%). They had a median age of 39 years (inter-quantile range (IQR) 31-46) and median disease duration of 7 years (IQR 0-12). mRNA isolation, labeling and hybridization was done at Expression Analysis (Q^2 Solutions, USA) in two batches. Labeled RNA was hybridized on Human Genome U133 Plus 2.0 Arrays (Affymetrix, USA). Sample data was processed using the R/Bioconductor *gcrma* library[53] and outliers were removed based on array quality scores. Expression data were normalized for technical factors—RNA quantity, quality (RIN score), and gene-level degradation slopes—and for batch effects, controlling for primary sample groups using the R *ComBat* library.[54]

Differential Expression (DE) analysis

Probe-level expression data were summarized to gene-level data using the *collapseRows* function from the R/WGCNA library.[55] Gene-level DE analysis was performed using the R *limma* library.[49] Prior to the analysis, the data were adjusted for 48 surrogate variables[54] to adjust for latent variability in the data, controlling for disease status. For individual linear regression, the R function *lm* was used to regress gene expression levels of individual genes on disease status.

Acknowledgements

We would like to thank Norm Allaire, Bryan Innis and Jaya Goyal for whole blood gene expression sample data generation. We also acknowledge the Analytical Biomolecular Research Facility at the University of Newcastle for flow cytometry support.

References

1. IMSGC. Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls. Am J Hum Genet. 2013;92: 854–865. doi:10.1016/j.ajhg.2013.04.019
2. Sawcer S, Franklin RJM, Ban M. Multiple sclerosis genetics. Lancet Neurol. Elsevier; 2014;13: 700–9. doi:10.1016/S1474-4422(14)70041-9
3. Olsson T, Barcellos LF, Alfredsson L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. Nat Rev Neurol. 2016; doi:10.1038/nrneurol.2016.187
4. IMSGC, WTCCC2, Sawcer S, Hellenthal G, Pirinen M, Spencer CCA, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011;476: 214–9. doi:10.1038/nature10251
5. IMSGC, Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet. 2013;45: 1353–60. doi:10.1038/ng.2770
6. IMSGC. The Multiple Sclerosis Genomic Map: Role of peripheral immune cells and resident microglia in susceptibility. bioRxiv. 2017; 1–43.
7. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523: 212–216. doi:10.1038/nature14465
8. Kulis M, Esteller M. DNA methylation and cancer. Adv Genet. 2010;70: 1–23. doi:10.1007/978-3-7643-8989-5_1
9. Long H, Yin H, Wang L, Gershwin ME, Lu Q. The critical role of epigenetics in systemic lupus erythematosus and autoimmunity. J Autoimmun. 2016;74: 118–138. doi:10.1016/j.jaut.2016.06.020
10. Ammal Kaidery N, Tarannum S, Thomas B. Epigenetic Landscape of Parkinson's Disease: Emerging Role in Disease Mechanisms and Therapeutic Modalities. Neurotherapeutics. 2013;10: 698–708. doi:10.1007/s13311-013-0211-8
11. International Parkinson's Disease Genomics Consortium (IPDGC), Wellcome Trust Case Control Consortium 2 (WTCCC2). A Two-Stage Meta-Analysis identifies several new loci for Parkinson's Disease. Gibson G, editor. PLoS Genet. Public Library of Science; 2011;7: e1002142. doi:10.1371/journal.pgen.1002142
12. Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P, Khrebtukova I, Miller N a, et al. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature. Nature Publishing Group; 2010;464: 1351–6. doi:10.1038/nature08990
13. Graves M, Benton M, Lea R, Boyle M, Tajouri L, Macartney-Coxson D, et al. Methylation differences at the HLA-DRB1 locus in CD4+ T-Cells are associated with multiple sclerosis. Mult Scler. 2014;20: 1033–1041. doi:10.1177/1352458513516529
14. Janson PCJ, Linton LB, Ahlen Bergman E, Marits P, Eberhardson M, Piehl F, et al. Profiling of CD4+ T Cells with Epigenetic Immune Lineage Analysis. J Immunol. 2011;186: 92–102. doi:10.4049/jimmunol.1000960

15. Maltby VE, Graves MC, Lea RA, Benton MC, Sanders KA, Tajouri L, et al. Genome-wide DNA methylation profiling of CD8+ T cells shows a distinct epigenetic signature to CD4+ T cells in multiple sclerosis patients. Clin Epigenetics. Clinical Epigenetics; 2015;7: 118. doi:10.1186/s13148-015-0152-7
16. Bos SD, Page CM, Andreassen BK, Elboudwarej E, Gustavsen MW, Briggs F, et al. Genome-Wide DNA Methylation Profiles Indicate CD8+ T Cell Hypermethylation in Multiple Sclerosis. PLoS One. 2015; 1–16. doi:10.7910/DVN/27694.Funding
17. Neven KY, Piola M, Angelici L, Cortini F, Fenoglio C, Galimberti D, et al. Repetitive element hypermethylation in multiple sclerosis patients. BMC Genet. BMC Genetics; 2016;17: 84. doi:10.1186/s12863-016-0395-0
18. Calabrese R, Zampieri M, Mechelli R, Annibali V, Guastafierro T, Ciccarone F, et al. Methylation-dependent *PAD2* upregulation in multiple sclerosis peripheral blood. Mult Scler J. 2012;18: 299–304. doi:10.1177/1352458511421055
19. Handel AE, De Luca GC, Morahan J, Handunnetthi L, Sadovnick AD, Ebers GC, et al. No evidence for an effect of DNA methylation on multiple sclerosis severity at HLA-DRB1*15 or HLA-DRB5. J Neuroimmunol. Elsevier B.V.; 2010;223: 120–123. doi:10.1016/j.jneuroim.2010.03.002
20. Maltby VE, Lea RA, Sanders KA, White N, Benton MC, Scott RJ, et al. Differential methylation at MHC in CD4 + T cells is associated with multiple sclerosis independently of HLA-DRB1. Clin Epigenetics. Clinical Epigenetics; 2017;9: 1–6. doi:10.1186/s13148-017-0371-1
21. Lin X, Barton S, Holbrook JD. How to make DNA methylome wide association studies more powerful. Epigenomics. 2016; doi:10.2217/epi-2016-0017
22. Quarles RH. Glycoproteins of myelin sheaths. J Mol Neurosci. 1997;8: 1–12. doi:10.1007/BF02736858
23. Mavrommatis E, Fish EN, Platanias LC. The Schlafen Family of Proteins and Their Regulation by Interferons. J Interf Cytokine Res. 2013;33: 206–210. doi:10.1089/jir.2012.0133
24. Puck A, Aigner R, Modak M, Cejka P, Blaas D, Stöckl J. Expression and regulation of Schlafen (SLFN) family members in primary human monocytes, monocyte-derived dendritic cells and T cells. Results Immunol. Elsevier; 2015;5: 23–32. doi:10.1016/j.rinim.2015.10.001
25. Chitnis T. The Role of CD4 T Cells in the Pathogenesis of Multiple Sclerosis. Int Rev Neurobiol. 2007;79: 43–72. doi:10.1016/S0074-7742(07)79003-7
26. Kaskow BJ, Baecher-Allan C. Effector T Cells in Multiple Sclerosis. Cold Spring Harb Perspect Med. 2018; a029025. doi:10.1101/cshperspect.a029025
27. Huseby ES, Huseby PG, Shah S, Smith R, Stadinski BD. Pathogenic CD8T cells in multiple sclerosis and its experimental models. Front Immunol. 2012;3: 1–9. doi:10.3389/fimmu.2012.00064
28. Nakagawa K, Matsuki T, Zhao L, Kuniyoshi K, Tanaka H, Ebina I, et al. Schlafen-8 is essential for lymphatic endothelial cell activation in experimental autoimmune encephalomyelitis. Int Immunol. 2018;30: 69–78. doi:10.1093/intimm/dxx079
29. Putnik M, Brodin D, Wojdacz TK, Fagerström-Billai F, Dahlman-Wright K, Wallberg AE. The transcriptional coregulator MAML1 affects DNA methylation and gene expression patterns in human embryonic kidney cells. Mol Biol Rep. 2016;43: 141–

150. doi:10.1007/s11033-016-3946-6
30. North ML, Jones MJ, MacIsaac JL, Morin AM, Steacy LM, Gregor A, et al. Blood and nasal epigenetics correlate with allergic rhinitis symptom development in the environmental exposure unit. Allergy. 2018;73: 196–205. doi:10.1111/all.13263
31. Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, et al. The UCSC Genome Browser database: 2018 update. Nucleic Acids Res. Oxford University Press; 2018;46: D762–D769. doi:10.1093/nar/gkx1020
32. The ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247
33. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13: 484–92. Available: http://www.ncbi.nlm.nih.gov/pubmed/22641018
34. Teschler S, Bartkuhn M, Künzel N, Schmidt C, Kiehl S, Dammann G, et al. Aberrant methylation of gene associated CpG sites occurs in borderline personality disorder. PLoS One. 2013;8: 1–10. doi:10.1371/journal.pone.0084180
35. Sun Z, Cunningham J, Slager S, Kocher J-P. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. Epigenomics. 2015;7: 813–28. doi:10.2217/epi.15.21
36. Bojesen SE, Timpson N, Relton C, Smith GD, Nordestgaard BG. AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. Thorax. 2017; no pagination. doi:10.1136/thoraxjnl-2016-208789
37. Marabita F, Almgren M, Sjöholm LK, Kular L, Liu Y, James T, et al. Smoking induces DNA methylation changes in Multiple Sclerosis patients with exposure-response relationship. Sci Rep. 2017;7: 14589. doi:10.1038/s41598-017-14788-w
38. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541: 81–86. doi:10.1038/nature20784
39. Fetahu IS, Höbaus J, Kállay E. Vitamin D and the epigenome. Front Physiol. 2014;5 APR: 1–12. doi:10.3389/fphys.2014.00164
40. Birdwell CE, Queen KJ, Kilgore PCSR, Rollyson P, Trutschl M, Cvek U, et al. Genome-Wide DNA Methylation as an Epigenetic Consequence of Epstein-Barr Virus Infection of Immortalized Keratinocytes. J Virol. 2014;88: 11442–11458. doi:10.1128/JVI.00972-14
41. Polman CH, Reingold SC, Banwell B, Clanet M, Cohen JA, Filippi M, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann Neurol. 2011;69: 292–302. doi:10.1002/ana.22366
42. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–75. doi:10.1086/519795
43. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34: 816–834. doi:10.1002/gepi.20533
44. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48: 1284–1287.

doi:10.1038/ng.3656
45. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30: 1363–9. doi:10.1093/bioinformatics/btu049
46. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. Genomics Data. 2016. pp. 22–24. doi:10.1016/j.gdata.2016.05.012
47. Jaffe AE, Hyde T, Kleinman J, Weinbergern DR, Chenoweth JG, McKay RD, et al. Practical impacts of genomic data “cleaning” on biological discovery using surrogate variable analysis. BMC Bioinformatics. BMC Bioinformatics; 2015;16: 372. doi:10.1186/s12859-015-0808-5
48. Leek JT. Asymptotic Conditional Singular Value Decomposition for High-Dimensional Genomic Data. Biometrics. 2011;67: 344–352. doi:10.1111/j.1541-0420.2010.01455.x
49. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47. doi:10.1093/nar/gkv007
50. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol. 2012;41: 200–209.
51. Fox RJ, Miller DH, Phillips JT, Hutchinson M, Havrdova E, Kita M, et al. Placebo-Controlled Phase 3 Study of Oral BG-12 or Glatiramer in Multiple Sclerosis. N Engl J Med. 2012;367: 1087–1097. doi:10.1056/NEJMoa1206328
52. Gold R, Kappos L, Arnold DL, Bar-Or A, Giovannoni G, Selmaj K, et al. Placebo-Controlled Phase 3 Study of Oral BG-12 for Relapsing Multiple Sclerosis. N Engl J Med. 2012;367: 1098–1107. doi:10.1056/NEJMoa1114287
53. Wu J, Irizarry R, MacDonald J, Gentry J. gcrma: Background Adjustment Using Sequence Information. R package.
54. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28: 882–883. doi:10.1093/bioinformatics/bts034
55. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, et al. Strategies for aggregating gene expression data: The collapseRows R function. BMC Bioinformatics. 2011;12: 322. doi:10.1186/1471-2105-12-322

Table 1. Characteristics of relapsing-remitting MS cases and controls, and count of CD4+ and CD8+ T cell samples included in analyses.

	Cases	Controls
Norway		
N	46	46
Age ± SD	38 ± 9	37 ± 9
Female	46 (100%)	46 (100%)
Treatment naïve[a]	44 (96%)	-
On treatment	2 (4%)	-
CD4+ T cell samples available for analysis	46 (100%)	41 (89%)
CD8+ T cell samples available for analysis	46 (100%)	46 (100%)
Australia		
N	48	53
Age ± SD	40 ± 11	45 ± 16
Female	48 (100%)	53 (100%)
Treatment naïve	16 (33%)	-
> 3 months off treatment	12 (25%)	-
On treatment[b]	22 (42%)	-
CD4+ T cell samples available for analysis	48 (100%)	53 (100%)
CD8+ T cell samples available for analysis	22 (46%)	11 (21%)

[a]Three individuals in this group were previously on treatment (5, 4, and 2.5 years before inclusion).
[b]One individual in this group had unknown treatment status.

Table 2: Overview of number of samples used in each of the five analyses, with quality control (QC) and analysis metrics.

T Cell Type	Cases, N	Controls, N	Probes Passing QC	SVs	λ	Candidate Regions	M-value Cutoff
All cases, regardless of treatment							
CD4+	94	94	423,500	13	1.11	3,989	0.154
CD8+	68	57	415,676	10	0.95	3,564	0.217
Treatment-naïve or off treatment for at least 3 months							
CD4+	72	94	423,500	12	1.10	3,902	0.170
Treatment-naïve							
CD4+	60	94	423,500	11	1.18	4,271	0.178
CD8+	44	46	409,357	7	1.02	3,703	0.198

Columns indicate the number of case and control samples included in each analysis, the count of CpG probes passing QC filters, the number of estimated surrogate variables (SVs), genomic inflation factor λ with SVs as covariates, the number of candidate differentially methylated regions tested, and the M-value cutoff determined by the *Bumphunter* R package.

Table 3. Differentially methylated regions (DMRs) in $CD4^+$ and $CD8^+$ T cells between MS cases and controls.

DMR Genomic Position (hg19)	DMR Position Relative to Genes	# CpGs in DMR	Direction of Methylation Change in Cases	Patient Subsets with DMR in CD4+ Cells	Patient Subsets with DMR in CD8+ Cells
chr6:29648225-29649084	8kb downstream of *MOG*; 3kb upstream of *ZFP57*	18-22	Decreased	All patients	-
				Treatment-naïve only	-
chr6:32551749-32552453	Exon 2 of *HLA-DRB1*	7-8	Decreased	All patients	All patients
				-	-
chr12:739980-740338	Intron of *NINJ2;* first exon of *LOC100049716*	3	Increased	-	-
				Treatment-naïve only	-
chr17:33734664-33734664	3kb downstream of *SLFN12*	1	Decreased	-	-
				Treatment-naïve only	-
chr17:33759512-33760527	First exon *SLFN12*	11-12	Increased	All patients	All patients
				Treatment-naïve only	Treatment-naïve only

P-values were adjusted for multiple tests, controlling the family-wise error rate (FWER from *Bumphunter*), and a DMR was called if the FWER was less than 0.2. Dash (-) indicates an FWER>0.2.

Table 4: Summary statistics for case/control expression differences in whole blood for genes identified in DNA methylation analyses.

Gene	FDR (*limma*)	p-value (*limma*)	p-value (linear fit)	Moderated log fold change (*limma*)	Log fold change (linear fit)
SLFN12	3.29e-09	2.25e-10	4.88e-11	-0.417	-0.599
HLA-DRB1	7.98e-12	3.17e-13	1.28e-13	-0.302	-0.442
NINJ2	1.97e-05	3.64e-06	1.83e-09	-0.186	-0.349
LOC100049716	0.527	0.446	0.046	-0.038	-0.123

Genome-wide analysis (*limma*) and individual linear fit coefficients and p-values are listed, as well as *limma* p-values adjusted for multiple hypothesis testing to control the false discovery rate (FDR). *Limma* moderates fold changes and hence represent more conservative coefficients and p-values than linear fit.

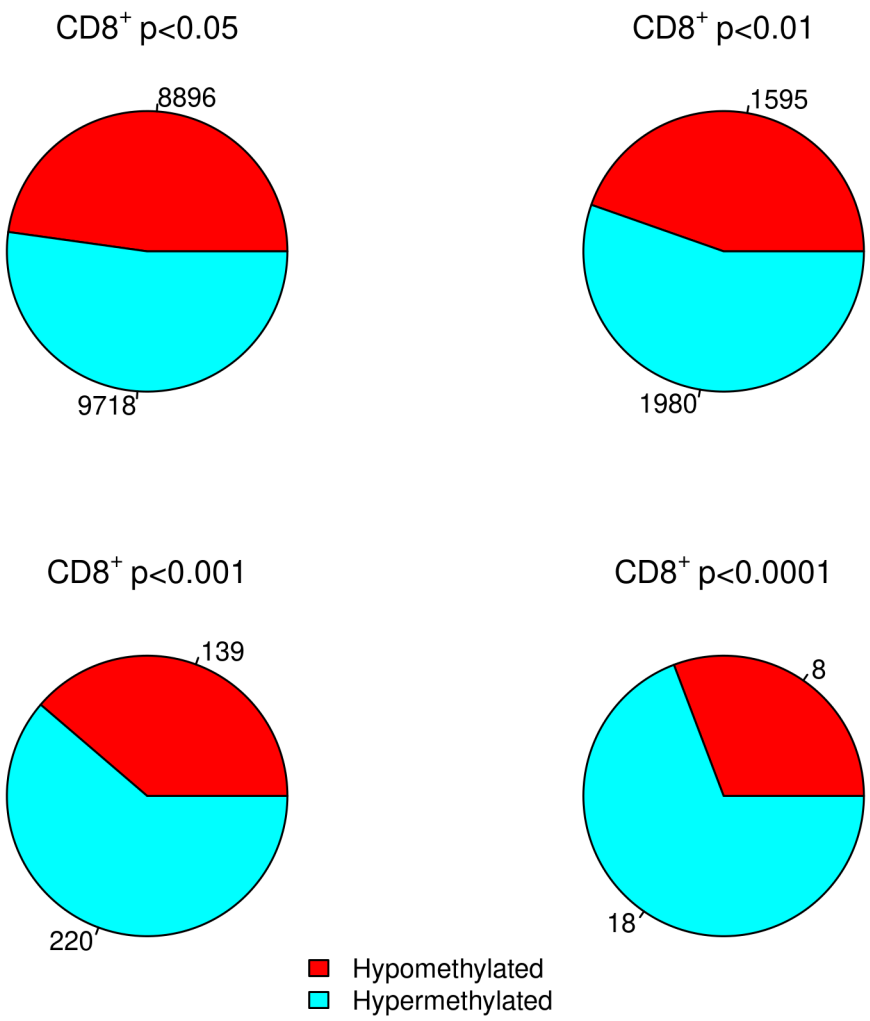


Fig 1. Proportion of significantly differentially methylated positions at increasingly stringent p-value cutoffs in the CD8+ T cells of 94 MS cases and 94 healthy controls. Numbers indicate the number of CpGs meeting the p-value threshold for hypomethylated and hypermethylated.

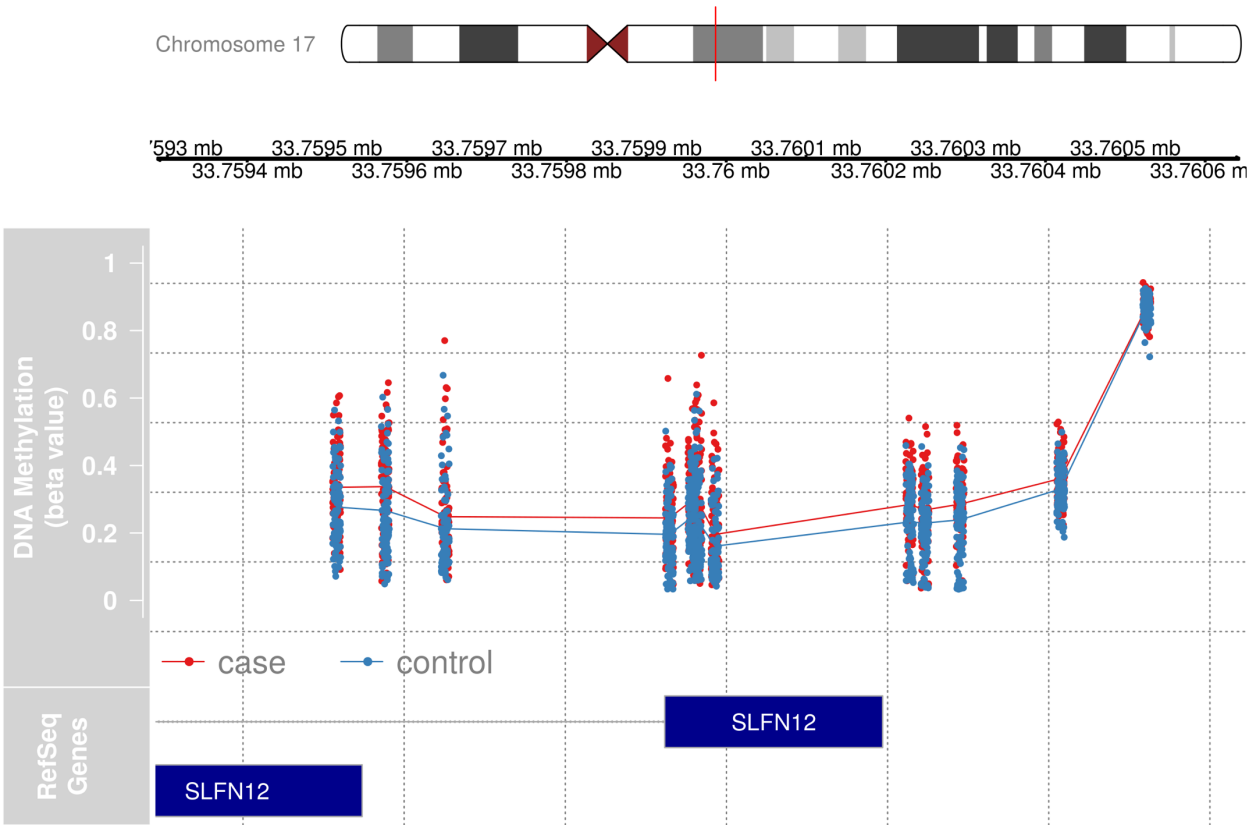


Fig 2. Detailed view of the differentially methylated region on chromosome 17, overlapping the first exon of *SLFN12*. Individual CpG sites and sample values from CD4$^+$ cells from 94 cases and 94 healthy controls are represented by dots (red dots – cases, blue dots – controls), whereas the lines represent the average values on each CpG site. The position of two *SLFN12* gene transcripts are shown in dark blue. *Illustration: the gviz package for R.*

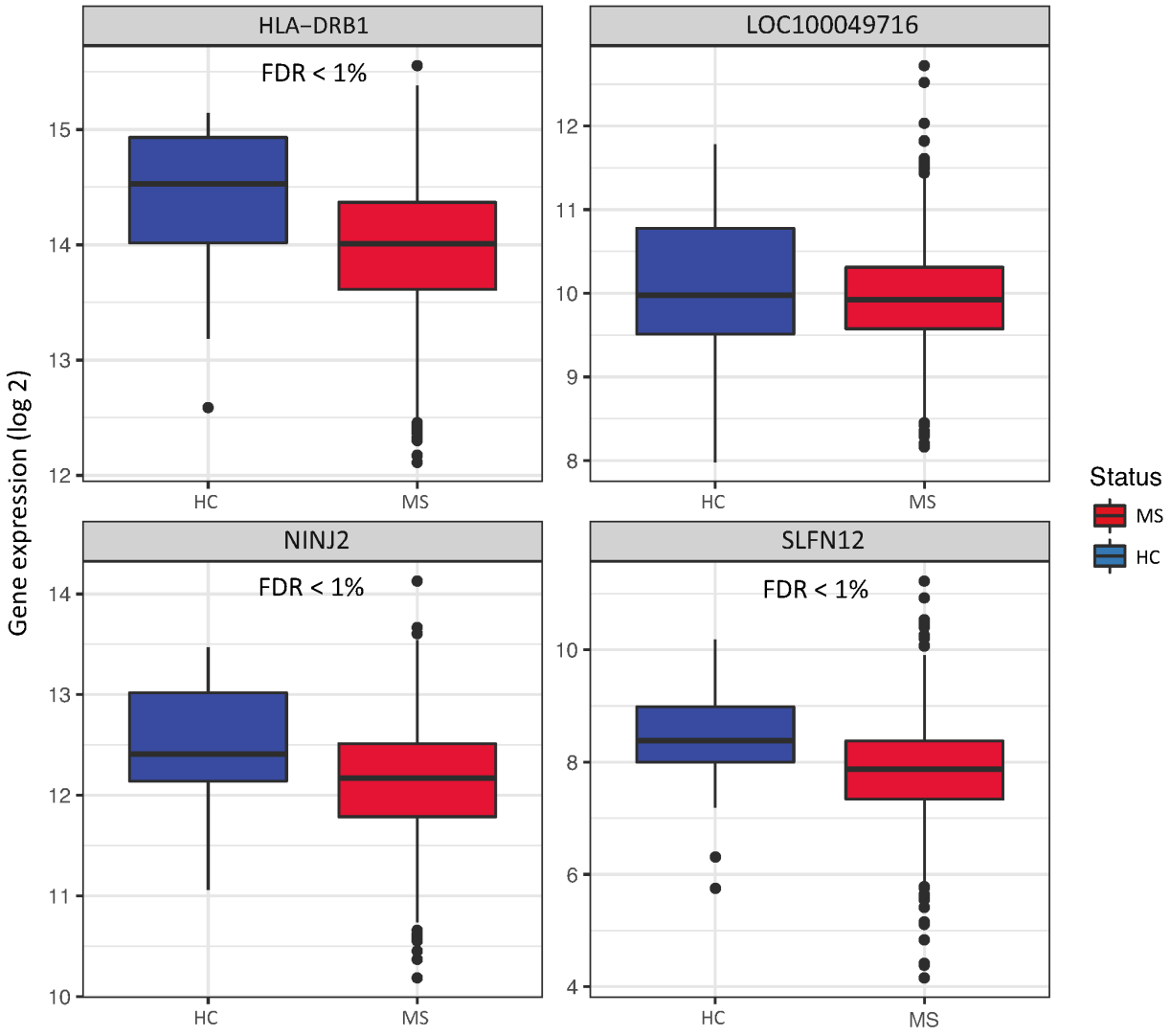


Fig 3. Gene expression levels of *HLA-DRB1, LOC100049716, NINJ2,* and *SLFN12* in whole blood of MS cases compared to healthy controls (HC). Horizontal lines of boxplots indicate the lower quartile, median, and upper quartile of log2 gene expression intensities; whiskers indicate the lowest and highest values within 1.5 times the inter-quartile range of the lower and upper quartiles; dots indicate outliers.

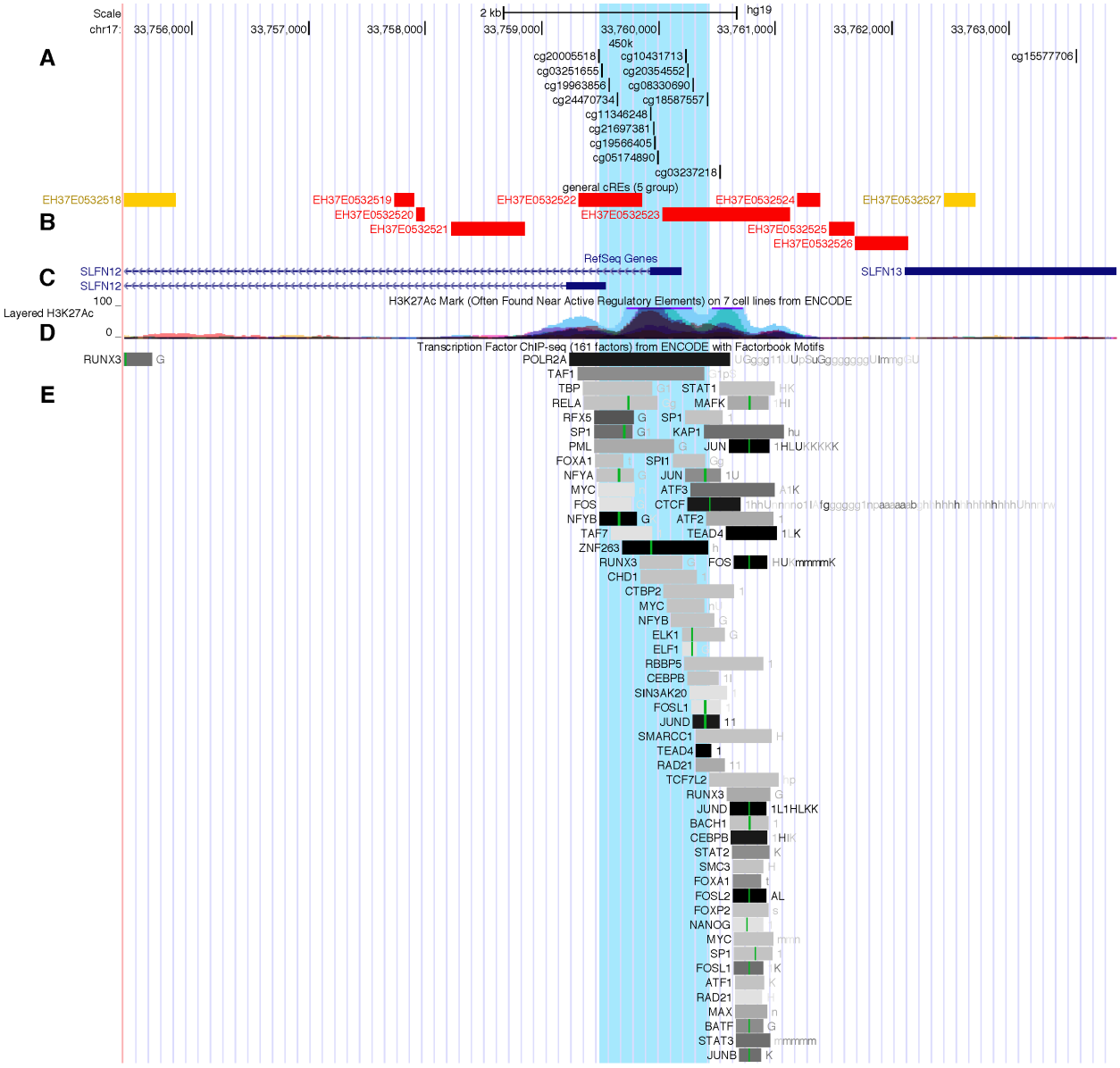
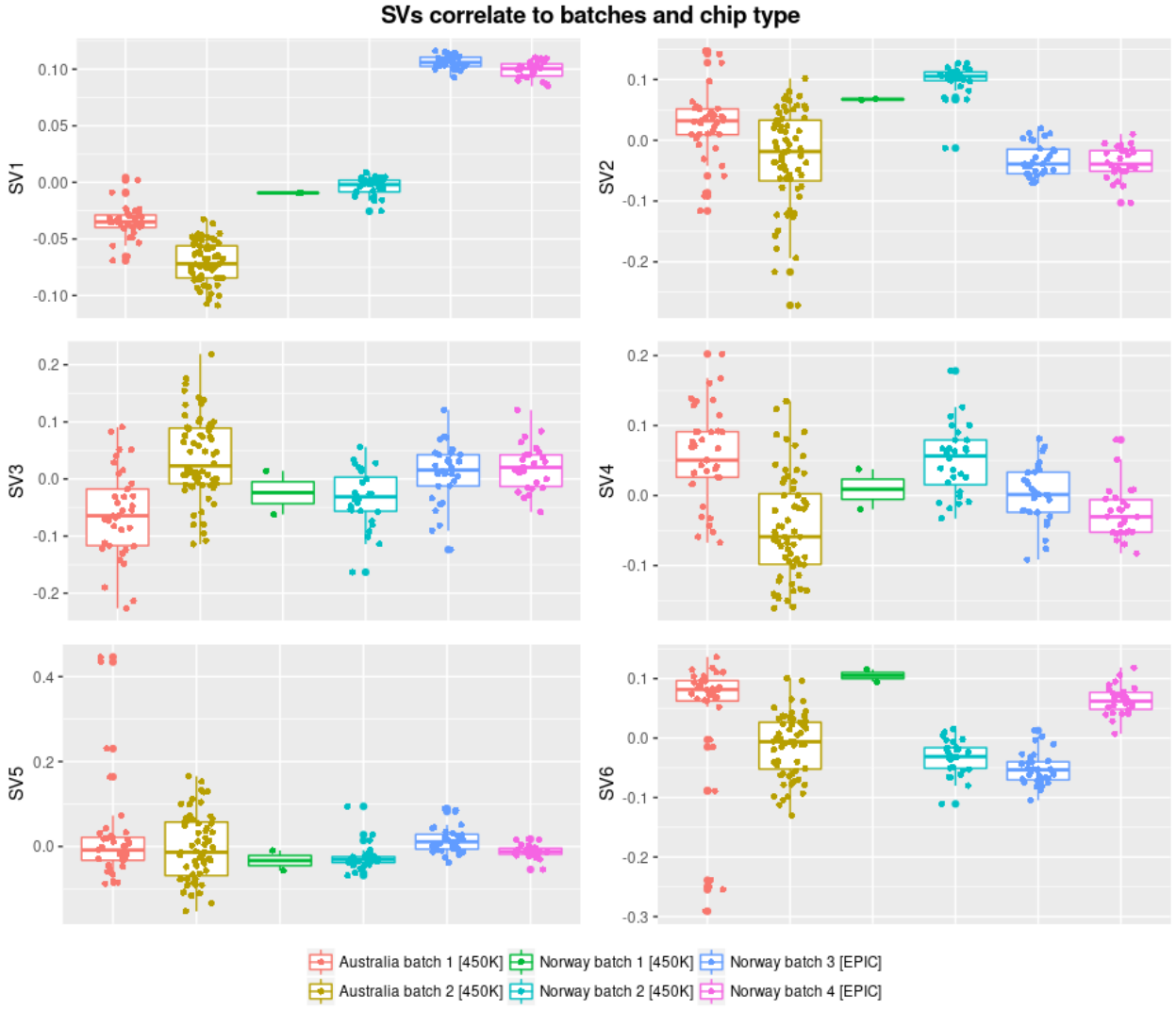
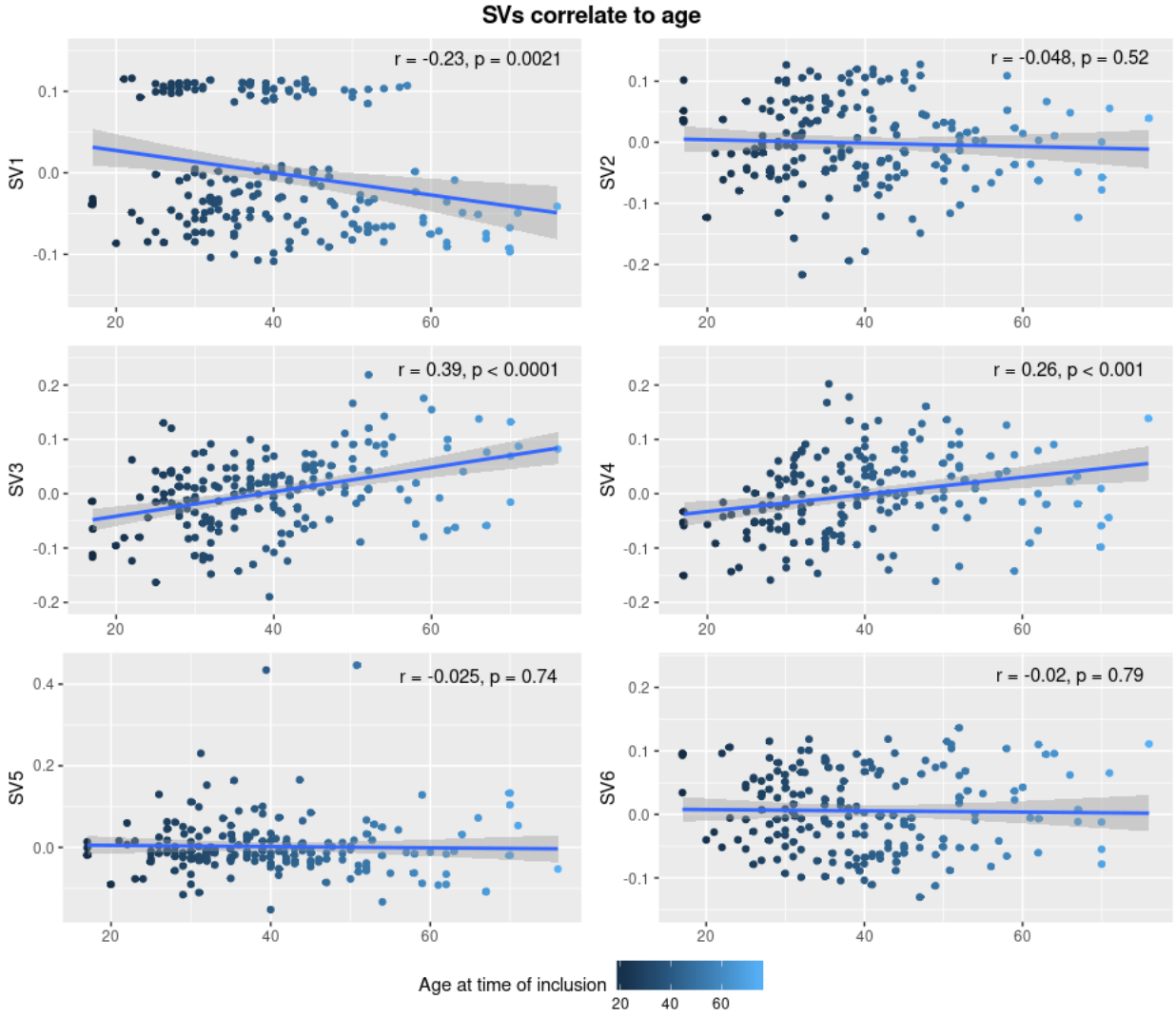


Fig 4. UCSC Genome Browser image showing the DMR in the *SLFN12* gene region highlighted in blue. Several tracks are shown: (A) the CpG sites on the Illumina 450k chip, (B) candidate regulatory elements identified by the ENCODE SCREEN algorithm, (C) the position of the *SLFN12* gene (the first exon of two transcript variants are shown), (D) the presence of the H3K27Ac mark across seven cell lines, and (E) the location of several transcription factor binding motifs in the genome sequence.

Supporting Information



S1 Fig. Box plots of the first 6 surrogate variables (SV1-SV6) from the CD4+ T cell analysis of all participants according to batch. Batch is correlated with each of the first 6 SVs except SV5. Illumina chip type (450k vs. EPIC) appears to be captured particularly well by SV1.



S2 Fig. Scatterplots of the first 6 surrogate variables (SV1-SV6) from the CD4+ T cell analysis of all participants according to participant age at the time of blood draw. Pearson correlation coefficients and p-values are given. SV3 and SV4 appear to capture age the best.

Chapter 4 - miRNA contributions to pediatric-onset multiple sclerosis inferred from GWAS

Abstract

Objective: Onset of multiple sclerosis (MS) occurs in childhood for approximately 5% of cases (pediatric MS, or ped-MS). Epigenetic influences are strongly implicated in MS pathogenesis in adults, including the contribution from microRNAs (miRNAs), small non-coding RNAs that affect gene expression by binding target gene mRNAs. Few studies have specifically examined miRNAs in ped-MS, but individuals developing MS at an early age may carry a relatively high burden of genetic risk factors, and miRNA dysregulation may therefore play a larger role in the development of ped-MS than in adult-onset MS. This study aimed to look for evidence of miRNA involvement in ped-MS pathogenesis.

Methods: GWAS results from 486 ped-MS cases and 1,362 controls from the U.S. Pediatric MS Network and Kaiser Permanente Northern California membership were investigated for miRNA-specific signals. First, enrichment of miRNA-target gene network signals was evaluated using MIGWAS software. Second, SNPs in miRNA genes and in target gene binding sites (miR-SNPs) were tested for association with ped-MS, and pathway analysis was performed on associated target genes.

Results: MIGWAS analysis showed that miRNA-target gene signals were enriched in GWAS ($p=0.038$) and identified 39 candidate biomarker miRNA-target gene pairs, including immune and neuronal signaling genes. The miR-SNP analysis implicated dysregulation of miRNA binding to target genes in 5 pathways, mainly involved in immune signaling.

Interpretation: Evidence from GWAS suggests that miRNAs play a role in ped-MS pathogenesis by affecting immune signaling and other pathways. Candidate biomarker miRNA-target gene pairs should be further studied for diagnostic, prognostic, and/or therapeutic utility.

Introduction

Multiple sclerosis (MS) is an immune-mediated demyelinating disease of the central nervous system and a leading cause of neurological disability in young adults. MS is typically diagnosed between the ages of 20 and 40, but it is estimated that up to 5% of all cases experience their first symptoms before the age of 18.[1,2] While pediatric-onset MS (ped-MS) and adult MS presentation largely overlap, disease course in children is almost exclusively relapsing-remitting, with a higher relapse rate, and a longer time to development of secondary progressive MS and disability.[2–4]

MS is thought to result from a complex interplay of genetic, epigenetic, and environmental risk factors.[5] MicroRNAs (miRNAs) are epigenetic factors that have been investigated in

MS, and over 170 miRNAs have been found to be differentially expressed in various tissues in either adult-onset MS or experimental autoimmune encephalomyelitis (EAE) in mice.[6–13] One study specifically compared miRNA expression levels in ped-MS cases to pediatric controls, and 12 upregulated and one downregulated miRNA were reported.[14] miRNAs are short (~22 nucleotides) non-coding RNAs that usually downregulate gene expression by binding to specific sequences of messenger RNA (mRNA) transcripts, targeting them for degradation and blocking protein translation, though different miRNA functions have also been reported.[15] Target sites generally lie in the 3' untranslated regions (3' UTRs) of mRNAs, but binding in other regions is known to occur.[16] Because each miRNA can target hundreds of genes, and any gene can be regulated by multiple miRNAs, they have the potential to influence entire networks of genes at once.

Single nucleotide polymorphisms (SNPs) in and around miRNA genes have been associated with a number of autoimmune diseases, including MS.[17] These miRNA SNPs can disrupt normal gene regulatory functions by affecting miRNA expression levels and processing, but SNPs in the target binding sites of mRNAs can also impact the normal function of miRNAs. Strategies to examine target gene SNPs in addition to miRNA gene SNPs have recently been developed and have implicated specific miRNAs and target genes in the development of autoimmune and other diseases.[18–21]

This study examined evidence of miRNA involvement in ped-MS susceptibility in two ways. First, genome-wide association study (GWAS) results were tested for enrichment of signals in miRNA-target gene networks utilizing MIGWAS software.[21,22] Second, a miR-SNP association study was performed, and pathway analysis was used to characterize target genes harboring miR-SNPs associated with ped-MS. We define miR-SNP as a SNP that is either (1) located in a gene that codes for a miRNA, or (2) located in a miRNA binding site of a target gene.

Methods

Study Participants

The study participants have been previously described,[23] but now include additional ped-MS cases with same inclusion/exclusion criteria. Additionally, only participants who were genotyped using an Illumina BeadChip array were included in this study. Briefly, patients with onset of MS or clinically isolated syndrome (CIS) with 2 silent MRI lesions suggestive of early MS before the age of 18 (n=432) were enrolled through the U.S. Network of Pediatric MS Centers.[3] Additional adult cases with onset prior to age 18 (n=68) were recruited retrospectively from the Kaiser Permanente Northern California (KPNC) membership. All cases were confirmed to have MS using established diagnostic criteria.[24,25]

Pediatric controls (n=208) were enrolled through the U.S. Network of Pediatric MS Centers.[3] Adult controls (n=894) without a diagnosis of MS, optic neuritis, transverse myelitis, or demyelinating disease confirmed through electronic medical records were recruited from KPNC and enrolled. A second set of previously described[26] female adult

controls (n=268) with no prior history of autoimmune disease who were recruited as part of the University of California San Francisco Mother-Child Immunogenetic Study were also included.

Data Collection

DNA from participants was purified from either whole blood or saliva samples. Genotyping was performed using Illumina Infinium 660K and Human Omni Express BeadChip arrays. Genotype data were merged into a single dataset and processed using PLINK 1.9.[27] SNPs with a minor allele frequency (MAF) <1% or success rate <90% and samples with >10% failed genotype calls were removed from analysis. To reduce confounding due to population stratification, analyses were restricted to white individuals, defined as having >=80% European ancestry identified using SNP weights for European, West African, East Asian and Native American ancestral populations.[28] Related individuals were identified, and only one randomly chosen person in any related group was retained for analysis, with the exception that cases were preferentially retained in instances where cases and controls were related. Classical multidimensional scaling (MDS) was used to visualize study population ancestry (Supplementary Figure) and include as covariates in subsequent statistical analyses. Five outlier samples were removed from analysis. SNP imputation was performed with reference haplotypes from Phase 3 of the 1000 Genomes Project[29] using SHAPEIT2 and IMPUTE2.[30] Genotypes were called using the default hard-call threshold of 90% using PLINK. Imputed SNPs with info score <0.3, with MAF <1%, with genotype call rate <90%, or not in Hardy-Weinberg equilibrium (HWE) among controls ($p<0.00001$) were removed. The final imputed dataset consisted of 7,570,644 autosomal SNPs, of which 42,277 lay within the major histocompatibility complex (MHC) region chr6:29570005-33377657 in GRCh37/hg19, spanning genes *GABBR1* to *KIFC1*.[31] Presence of the *HLA-DRB1*15:01* allele, the strongest genetic risk factor for MS, was determined for each participant using the tag SNP rs3135388.[32] There were 486 cases and 1,362 controls in the final dataset. Study protocols were approved by all institutions, and informed consent or assent was obtained from all participants, as previously described.[23]

Statistical Analyses - miRNA-Target Gene Network Enrichment in GWAS (MIGWAS)

Genome-wide association tests were performed on all autosomal SNPs outside of the MHC using logistic regression and additive genetic models in PLINK. Models included the first 3 MDS components to adjust for residual confounding by population stratification. Enrichment of miRNA-target gene network signals in the GWAS results was evaluated using MIGWAS software.[21,22] Briefly, MIGWAS takes GWAS p-values as input, selects the lowest p-value per miRNA and target gene, and, for each of 179 different tissues with available miRNA expression data from FANTOM5,[33] identifies *the number of miRNA-target gene pairs* that satisfy the following conditions: both the miRNA and the target gene are associated with the outcome ($p<0.01$), there is a high binding score prediction for the pair, and the miRNA is highly and specifically expressed in the tissue. Enrichment of miRNA-target gene signal is estimated by permuting GWAS p-values 20,000 times and

recomputing the number of miRNA-target gene pairs that satisfy the conditions to obtain an empirical null distribution of that number. A p-value for enrichment is then reported for each tissue, as well as an overall enrichment p-value that does not take tissue expression into consideration is also reported. Enrichment p-values of 0.05 or lower were considered significant. Candidate biomarker miRNA-target gene pairs are also reported, and are defined as pairs where both the miRNA and target gene are nominally associated with the outcome ($p<0.01$) and the miRNA-target gene binding prediction score is in the top one percentile of all pairs.[22]

Statistical Analyses - miR-SNP Association

miR-SNPs were tested separately for association with ped-MS. miR-SNPs in miRNA genes were identified using version 21 high-confidence annotations from the miRBase database.[34] Coordinates for 1,877 miRNA genes were converted from build 38 to build 37 using the UCSC Genome Browser liftOver tool[35] and intersected with imputed SNPs in the ped-MS dataset using BEDTools,[36] resulting in 267 miR-SNPs outside of the MHC and 8 within the MHC. miR-SNPs in predicted target binding sites in the 3' UTRs of protein-coding genes were identified using the MirSNP database[18] and version 3.0 of the PolymiRTS database.[19] SNPs in predicted target regions from either database were intersected with the ped-MS dataset, resulting in 51,725 target-region miR-SNPs outside of the MHC and 586 miR-SNPs within the MHC.

Each miR-SNP was tested for association with ped-MS using the same logistic regression models as in the GWAS analysis, with the exception that models for miR-SNPs within the MHC also included the *HLA-DRB1*15:01* tag SNP as a covariate. P-values in each category of association tests—miRNA gene SNPs or target gene SNPs, within or outside of the MHC—were adjusted separately for multiple hypothesis testing using the Benjamini-Hochberg procedure to control the false discovery rate.[37] A threshold of 0.05 was used to determine significance. In an effort to reduce the statistical burden of multiple hypothesis tests, several sets of candidate mir-SNPs were considered separately: 897 miR-SNPs in the 3'UTRs of genes proximal to 200 non-MHC loci identified in the latest adult-onset MS genome-wide association study;[38] 1,063 miR-SNPs experimentally supported by crosslinking, ligation, and sequencing of hybrids (CLASH) experiments, including in non-canonical binding sites and non-protein-coding genes;[16] 19,516 miR-SNPs predicted in the polymiRTS database to either create a new miRNA binding site or disrupt a conserved miRNA binding site, with a context+ score difference of less than -0.15 (more negative scores indicate increased confidence that miRNA binding is disrupted); and one miR-SNP in the 3' UTR of the *HLA-DRB1* gene.

Pathway analysis of target genes with mir-SNPs that were nominally associated with ped-MS at $p<0.01$ in the main association analysis were conducted using PANTHER.[39] Statistical overrepresentation tests were performed using a background list of only protein-coding genes. Each of 9 available annotation data sets in the "PANTHER," "GO," and "Reactome" pathways were tested using Fisher's Exact test with FDR multiple test correction.

Results

Characteristics of study participants are summarized in Table 1. Average age of onset for ped-MS cases was 14.3 years, and cases had more copies of the *HLA-DRB1*15:01* allele than controls, as expected. When plotted with HGDP reference populations, cases and controls clustered together near European individuals (Supplementary Figure).

miRNA-Target Gene Network Enrichment in GWAS (MIGWAS)

Enrichment of miRNA-target gene network signals was observed in the ped-MS GWAS results for 25 different tissues ($p<0.05$) as well as overall, without considering tissue-specific miRNA expression ($p=0.038$). Results are summarized in Table 2. MIGWAS identified 39 candidate biomarker miRNA-target gene pairs comprised of 16 unique miRNAs and 37 unique genes (Table 3).

miR-SNP Association

After adjusting p-values for multiple hypothesis testing, no miR-SNPs were significantly associated with ped-MS in the genome-wide analyses at FDR<0.05. There were 255 target genes with 3' UTR miR-SNPs associated at $p<0.01$ in the genome-wide analysis that were used as input for pathway analyses. These genes were overrepresented in five pathways in PANTHER (Table 4). Among the candidate miR-SNP sets, only one CLASH-supported miR-SNP, rs61075345 in the third exon of *TVP23B*, was associated with ped-MS ($p=4.59 \times 10^{-05}$, FDR= 0.047).

Discussion

In the current study, evidence that miRNAs are involved in ped-MS pathogenesis was sought using two different approaches that utilized genetic data from the largest population of ped-MS cases gathered to date.

The MIGWAS method identified enrichment of miRNA-target gene networks in ped-MS GWAS results, and identified tissues in which miRNAs involved in those networks are known to be highly expressed. Tissues included gastrointestinal, brain, fetal, fat, joint, immune, lung, vascular, skin, kidney and other tissues (Table 2). While it is true that immune and central nervous system tissues have clear roles in MS and are generally prioritized first for study, there is evidence that processes starting in other tissues may play a role in triggering MS and also warrant investigation. For instance, smoking is hypothesized to exert its effect on MS risk primarily through irritation and inflammation of lung tissue, which in turn likely trigger (possibly auto-reactive) immune responses.[5] The highest enrichment observed in this ped-MS study was in a gastrointestinal tissue, *keratinized cells of the oral mucosa*, and gastrointestinal tissues were overrepresented in these results: 4 of the 7 gastrointestinal tissues tested were enriched for miRNA-target gene network signals. Evidence that miRNA dysregulation could specifically be occurring

in gastrointestinal tissues is notable because there is existing evidence of a bi-directional relationship between MS and the gut microbiome, where aberrant gut microbiomes found in MS patients contribute to a pro-inflammatory state, and the autoreactive immune systems of MS patients shape the gut microbiome.[40]

Several genes identified in the MIGWAS candidate biomarker target-gene pairs are involved in immune signaling and activation (Table 3) according to RefSeq annotations,[41] and are therefore particularly promising targets of future research into the role miRNAs play in ped-MS development. *CIITA* is a "master regulator" of class II HLA gene expression, and *CD80* is a T cell membrane receptor that provides the costimulatory signal necessary for T cell activation. *CD109* is expressed in activated T cells and regulates transforming growth factor beta signaling. *CBL* is an enzyme required for targeting substrates for degradation by the proteasome and is a negative regulator of many signaling pathways triggered by activation of cell surface receptors. *TFAP4* is a transcription factor that activates both viral and cellular genes. Two other MIGWAS genes with plausible roles in ped-MS pathogenesis affect neuronal differentiation and signaling. *GLIS2* is widely expressed at low levels in the neural tube and peripheral nervous system and is thought to promote neuronal differentiation, and *NCS1* modulates synaptic transmission and synaptic plasticity and is expressed predominantly in neurons. Three genes identified by MIGWAS are involved in protein folding and homeostasis in the endoplasmic reticulum (ER), which is notable because the ER lumen cellular component was also identified in the miR-SNP pathway analysis. The ER lumen is where class I and II HLA proteins are assembled,[42] and stress in the ER caused by accumulation of misfolded proteins (the "unfolded protein response" or UPR) is associated with a number of inflammatory diseases, including MS.[43] Of the MIGWAS-identified genes, *HYOU1* is thought to play an important role in protein folding and secretion in the ER. *ERP29* localizes to the lumen of the ER and is involved in the processing of secretory proteins. *SLC37A4* regulates transport from the cytoplasm to the lumen of the ER to maintain glucose homeostasis and plays a role in calcium sequestration in the ER lumen. Collectively, the genes in the 39 miRNA-target gene pairs suggest that miRNAs could be affecting ped-MS through many mechanisms, including immune signaling and activation, neuronal differentiation and signaling, and protein folding in the ER. Finally, it is notable that expression differences for five of the 16 candidate miRNAs identified by MIGWAS, hsa-mir-197, hsa-mir-200c, hsa-mir-21, hsa-mir-599, and hsa-mir-744, have been associated with MS or EAE in previous studies,[6,7] and two of them, hsa-miR-21 and hsa-miR-3605, were differentially expressed in ped-MS cases specifically (though hsa-miR-21 failed a subsequent validation assay).[14] A follow-up to the ped-MS expression study found that six of the 13 confirmed ped-MS-associated miRNAs were also differentially expressed in adults,[44] but hsa-miR-3605 was not among them, suggesting that it could be a biomarker specific to ped-MS.

In the miR-SNP analysis, the single CLASH-supported miR-SNP associated with ped-MS resides in the *TVP23B* gene, which codes for a membrane protein associated with diabetic retinopathy.[45] It is not immediately clear how it may play a role in ped-MS pathogenesis. However, statistical overrepresentation tests of top miR-SNP hits yielded

two receptor-mediated signaling pathways with a more evident relationship to ped-MS. Five genes in the histamine H_1 receptor pathway were found to have ped-MS-associated miR-SNPs (Table 4). Histamine is a ubiquitous compound in human tissues that acts as a neurotransmitter and that is involved in inflammatory responses that act through four different receptors, H_1-H_4. It is thought that the pro-inflammatory effects of histamine act through H_1 receptors.[46] Many of the same genes are also involved in the 5-HT_2 type receptor mediated signaling pathway. 5-HT_2 is a subtype of serotonin receptors. Similar to histamine, serotonin is a signaling molecule with wide-ranging effects that acts as both a neurotransmitter and a hormone. The 5-HT_2 class of hormone receptors is expressed on several immune cell types.[47] Our results suggest that dysregulation of genes involved in these signaling pathways by miRNAs increases ped-MS risk.

The other three pathways identified in the miR-SNP analysis each encompass many of the same genes, including several genes encoded in the MHC (Table 4). Five class I and II HLA genes associated with ped-MS are in the *MHC protein complex (GO cellular component)*. These genes code for proteins that present antigens to T cells, and variants in antigen-presenting genes are the first-documented and strongest genetic risk factors for MS.[5,38] The same five HLA genes are part of the *integral component of the luminal side of endoplasmic reticulum* (ER) *membrane (GO Cellular Component)*. Of note, the genes identified in the miR-SNP analysis are different from those identified in the MIGWAS analysis, but both methods point to dysregulation of processes in the ER. The *interferon gamma* (IFN-γ) *signaling* pathway contains a total of nine ped-MS-associated genes. The role of IFN-γ in MS has been extensively studied, and, similar to histamine and serotonin, it can have detrimental or beneficial effects on MS depending on where and when it is active.[48] Our miR-SNP analysis findings indicate that aberrant regulation by miRNAs of genes in the MHC protein complex, genes on the inner part of the ER lumen, or genes involved in IFN-γ signaling (which are not mutually exclusive genes), could be contributing to ped-MS pathogenesis.

This study had some limitations. It is possible that the SNPs identified in the MIGWAS and miR-SNP studies do not affect ped-MS via miRNA function but instead are associated due to linkage disequilibrium with SNPs acting by other mechanisms. Another issue is that miRNA-target binding prediction is imperfect, and therefore some of the miRNAs may not actually act on the genes identified in MIGWAS, and some of the miR-SNPs tested may not in reality impact miRNA function. Because the study was restricted to a white population, results may not be generalizable, and miRNA associations that exist in other non-white populations may have been missed.

An important strength of this study is that it utilized the largest study population thus far for ped-MS, which is a rare disease, and therefore difficult to study. Furthermore, cases were ascertained by a panel of pediatric MS specialists. Only samples genotyped on Illumina microarrays were utilized, minimizing the possibility of imputation bias, and rigorous quality control of microarray data was applied. By assessing p-values of SNPs in miRNA and target genes at the same time, and by including miRNA expression data,

MIGWAS was able to detect signals that may be missed with traditional GWAS or miR-SNP analysis.

In conclusion, this study provides evidence that ped-MS risk is influenced by miRNAs acting on immune signaling and other genes, and several miRNA-target gene pairs and specific tissues were nominated for further study. Larger studies are needed to confirm these results, and further work is needed to determine whether any miRNA-mediated disease processes are specific to the pediatric population.

Acknowledgements

The authors wish to thank Jorge Oksenberg for assisting with processing and DNA extraction of patient and control samples, Hans Christian von Buedingen for assisting with cell sorting for the miRNA analysis, and Shelly Roalstad for assisting with data collection. This work was supported in part by the NIH NINDS: 1R01NS071463 (PI: Waubant), R01NS049510 (PI: Barcellos), F31NS096885 (PI: Rhead); NIH NIEHS: R01ES017080 (PI: Barcellos), NIH NIAID: R01AI076544 (PI: Barcellos), the National MS Society HC 0165 (PI: Casper), and Race to Erase MS (PI: Waubant).

Conflicts of interest

E. Waubant is site PI for a Novartis and Roche trial. She has volunteered on an advisory board for a Novartis trial. She is a non-remunerated advisor for clinical trial design to Novartis, Biogen-IDEC, Sanofi, Genentech, Serono and Celgene. She has funding from the NMSS, PCORI and the Race to Erase MS. She is the section editor for Annals of Clinical and Translational Neurology, and co-Chief editor for MS And Related Disorders. A. Waldman reports grants from NIH (NINDS) NS071463 and from NMSS during the conduct of the study, as well as funds for investigator-initiated study from Ionis Pharmaceuticals and Biogen Idec and grants from United Leukodystrophy Foundation outside the submitted work. She is a consultant for Optum and has received royalties from UpToDate. B. Greenberg has received grant funding from Chugai, Medimmune, Medday, NIH, NMSS, Guthy Jackson Charitable Foundation, and Transverse Myelitis Association. He has received consulting fees from Novartis, EMD Serono, Celgene, and Alexion. L. Benson reports BG12 clinical trial support from Biogen, grants from Boston Children's Hospital Office of Faculty Development, travel funds from National MS Society, and personal fees from National Vaccine Compensation Program outside the submitted work. T.C. Casper reports grants from National MS Society. J. Graves reports speaking honoraria from Novartis outside the submitted work. B. Weinstock-Guttman reports grants and personal fees from Biogen, EMD Serono, Novartis, and Genentech, and personal fees from Mallinckrodt outside the submitted work.

References

1. Chitnis T, Glanz B, Jaffin S, Healy B. Demographics of pediatric-onset multiple sclerosis in an MS center population from the Northeastern United States. Mult. Scler. J. 2009;15(5):627–631.
2. Renoux C, Vukusic S, Mikaeloff Y, et al. Natural History of Multiple Sclerosis with Childhood Onset. N. Engl. J. Med. 2007;25(356):2603–2613.
3. Belman AL, Krupp LB, Olsen CS, et al. Characteristics of Children and Adolescents With Multiple Sclerosis. Pediatrics 2016;138(1)
4. Gorman MP, Healy BC, Polgar-Turcsanyi M, Chitnis T. Increased relapse rate in pediatric-onset compared with adult-onset multiple sclerosis. Arch. Neurol. 2009;66(1):54–59.
5. Olsson T, Barcellos LF, Alfredsson L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. Nat. Rev. Neurol. 2016;
6. Roopali Gandhi. miRNA in Multiple Sclerosis: Search for Novel Biomarkers. Mult. Scler. J. 2015;21(9):1095–1103.
7. Huang Q, Xiao B, Ma X, et al. MicroRNAs associated with the pathogenesis of multiple sclerosis. J. Neuroimmunol. 2016;295–296:148–161.
8. Yang Q, Pan W, Qian L. Identification of the miRNA–mRNA regulatory network in multiple sclerosis. Neurol. Res. 2017;39(2):142–151.
9. Groen K, Maltby VE, Lea RA, et al. Erythrocyte microRNA sequencing reveals differential expression in relapsing-remitting multiple sclerosis. BMC Med. Genomics 2018;11(1):1–12.
10. Regev K, Healy BC, Paul A, et al. Identification of MS-specific serum miRNAs in an international multicenter study. Neurol. - Neuroimmunol. Neuroinflammation 2018;5(5):e491.
11. Selmaj I, Cichalewska M, Namiecinska M, et al. Global exosome transcriptome profiling reveals biomarkers for multiple sclerosis. Ann. Neurol. 2017;81(5):703–717.
12. Teymoori-Rad M, Mozhgani SH, Zarei-Ghobadi M, et al. Integrational analysis of miRNAs data sets as a plausible missing linker between Epstein-Barr virus and vitamin D in relapsing remitting MS patients. Gene 2019;689(August 2018):1–10.
13. Venkatesha S, Dudics S, Song Y, et al. The miRNA Expression Profile of Experimental Autoimmune Encephalomyelitis Reveals Novel Potential Disease Biomarkers. Int. J. Mol. Sci. 2018;19(12):3990.
14. Liguori M, Nuzziello N, Licciulli F, et al. Combined microRNA and mRNA expression analysis in pediatric multiple sclerosis: An integrated approach to uncover novel pathogenic mechanisms of the disease. Hum. Mol. Genet. 2018;27(1):66–79.
15. Dragomir MP, Knutsen E, Calin GA. SnapShot: Unconventional miRNA Functions. Cell 2018;174(4):1038–1038.e1.
16. Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell 2013;153(3):654–665.
17. Latini A, Ciccacci C, Novelli G, Borgiani P. Polymorphisms in miRNA genes and their involvement in autoimmune diseases susceptibility. Immunol. Res. 2017;65(4):811–827.

18. Liu C, Zhang F, Li T, et al. MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. BMC Genomics 2012;13(1):661.
19. Bhattacharya A, Ziebarth JD, Cui Y. PolymiRTS Database 3.0: Linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. Nucleic Acids Res. 2014;42(D1):86–91.
20. de Almeida RC, Chagas VS, Castro MAA, Petzl-Erler ML. Integrative analysis identifies genetic variants associated with autoimmune diseases affecting putative microRNA binding sites. Front. Genet. 2018;9(APR):1–13.
21. Sakaue S, Hirata J, Maeda Y, et al. Integration of genetics and miRNA–target gene network identified disease biology implicated in tissue specificity. Nucleic Acids Res. 2018;1–12.
22. Okada Y, Muramatsu T, Suita N, et al. Significant impact of miRNA-target gene networks on genetics of human complex traits. Sci. Rep. 2016;6:1–9.
23. Gianfrancesco MA, Stridh P, Shao X, et al. Genetic risk factors for pediatric-onset multiple sclerosis. Mult. Scler. J. 2017;1–10.
24. Krupp LB, Tardieu M, Amato MP, et al. International Pediatric Multiple Sclerosis Study Group criteria for pediatric multiple sclerosis and immune-mediated central nervous system demyelinating disorders: revisions to the 2007 definitions. Mult. Scler. 2013;19(10):1261–7.
25. Polman CH, Reingold SC, Banwell B, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann. Neurol. 2011;69(2):292–302.
26. Cruz GI, Shao X, Quach H, et al. Increased risk of rheumatoid arthritis among mothers with children who carry *DRB1* risk-associated alleles. Ann. Rheum. Dis. 2017;1–6.
27. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007;81(3):559–75.
28. Chen CY, Pollack S, Hunter DJ, et al. Improved ancestry inference using weights from external reference panels. Bioinformatics 2013;29(11):1399–1406.
29. The 1000 Genomes Project Consortium, Auton A, Abecasis GR, et al. A global reference for human genetic variation. Nature 2015;526(7571):68–74.
30. Howie B, Fuchsberger C, Stephens M, et al. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 2012;44(8):955–9.
31. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction , diversity and disease. J. Hum. Genet. 2009;54(January):15–39.
32. de Bakker PIW, McVean G, Sabeti PC, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat. Genet. 2006;38(10):1166–72.
33. De Rie D, Abugessaisa I, Alam T, et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat. Biotechnol. 2017;35(9):872–878.
34. Kozomara A, Griffiths-Jones S. MiRBase: Annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42(D1):68–73.

35. Tyner C, Barber GP, Casper J, et al. The UCSC Genome Browser database: 2017 update. Nucleic Acids Res. 2017;45(D1):D626–D634.
36. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841–842.
37. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 1995;57(1):289–300.
38. IMSGC. The Multiple Sclerosis Genomic Map: Role of peripheral immune cells and resident microglia in susceptibility. bioRxiv 2017;1–43.
39. Mi H, Huang X, Muruganujan A, et al. PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017;45(D1):D183–D189.
40. Kirby T, Ochoa-Repáraz J. The Gut Microbiome in Multiple Sclerosis: A Potential Therapeutic Avenue. Med. Sci. 2018;6(3):69.
41. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–D745.
42. van Kasteren SI, Overkleeft H, Ovaa H, Neefjes J. Chemical biology of antigen presentation by MHC molecules. Curr. Opin. Immunol. 2014;26(1):21–31.
43. Stone S, Lin W. The unfolded protein response in multiple sclerosis. Front. Neurosci. 2015;9(JUL):1–11.
44. Nuzziello N, Vilardo L, Pelucchi P, et al. Investigating the Role of MicroRNA and Transcription Factor Co-regulatory Networks in Multiple Sclerosis Pathogenesis. Int. J. Mol. Sci. 2018;19(11):1–18.
45. Wang AL, Rao VR, Chen JJ, et al. Role of FAM18B in diabetic retinopathy. Mol. Vis. 2014;20(January 2012):1146–1159.
46. Jadidi-Niaragh F, Mirshafiey A. Histamine and histamine receptors in pathogenesis and treatment of multiple sclerosis. Neuropharmacology 2010;59(3):180–189.
47. Herr N, Bode C, Duerschmied D. The Effects of Serotonin in Immune Cells. Front. Cardiovasc. Med. 2017;4(July):1–11.
48. Arellano G, Ottum PA, Reyes LI, et al. Stage-specific role of interferon-gamma in experimental autoimmune encephalomyelitis and multiple sclerosis. Front. Immunol. 2015;6(SEP)

Tables and Figures

Table 1: Characteristics of ped-MS case and control individuals in the miR-SNP association study.

	Ped-MS Cases	**Controls**
N	486	1,362
Sex		
Female	362 (74)	1,122 (82)
Male	124 (26)	240 (18)
Age of onset	14.3 (3.2)	--
Copies *HLA-DRB1*15:01* allele		
0	250 (51)	1,005 (74)
1	194 (40)	334 (24)
2	42 (9)	23 (2)

Table values are mean (SD) for continuous variables or n (%) for categorical variables.

Table 2: Tissues enriched for miRNA-target gene network signals ($p<0.05$) in ped-MS GWAS results in the MIGWAS analysis.

Tissue	P-value	Fold change	MIGWAS tissue category
Keratinized cell of the oral mucosa	0.002	4.07	gastrointestinal
Human spinal cord - adult sample	0.011	2.61	brain
Epithelial cell of amnion	0.014	2.99	fetal
Preadipocyte	0.020	2.42	fat
Amnion mesenchymal stem cell	0.020	2.74	fetal
Epithelial cell of alimentary canal	0.023	2.79	gastrointestinal
Synovial cell	0.025	2.25	joint
Epithelial cell of esophagus	0.026	2.28	gastrointestinal
Acinar cell of sebaceous gland	0.027	2.64	fat
Mast cell	0.029	2.49	immune
Non-pigmented ciliary epithelial cell	0.031	2.13	skin
Tracheal epithelial cell	0.033	2.59	lung
Smooth muscle cell of internal thoracic artery	0.035	2.14	vascular
All (tissue-naïve test)	0.038	1.68	-
Extraembryonic cell	0.036	2.42	fetal
Human renal cortical epithelial cell sample	0.039	1.79	kidney
Pericyte cell	0.041	2.02	others
Hair follicle dermal papilla cell	0.041	2.46	skin
Mesangial cell	0.043	2.03	kidney
Keratinizing barrier epithelial cell	0.043	2.54	others
Gingival epithelial cell	0.043	1.86	gastrointestinal
CD14-pos CD16-neg classical monocyte	0.044	2.30	immune
Exocrine cell	0.045	2.42	others
Epidermal cell	0.046	2.37	others
Omentum preadipocyte	0.047	2.25	fat
Keratinocyte	0.050	2.43	skin

P-values and fold changes are for enrichment of the number of miRNA-target gene pairs associated with ped-MS (where the pair has a high predicted binding score and the miRNA is highly expressed in the tissue) compared to the empirical null distribution of the number of such pairs.

Table 3: Candidate biomarker ped-MS miRNA-target gene pairs from MIGWAS.

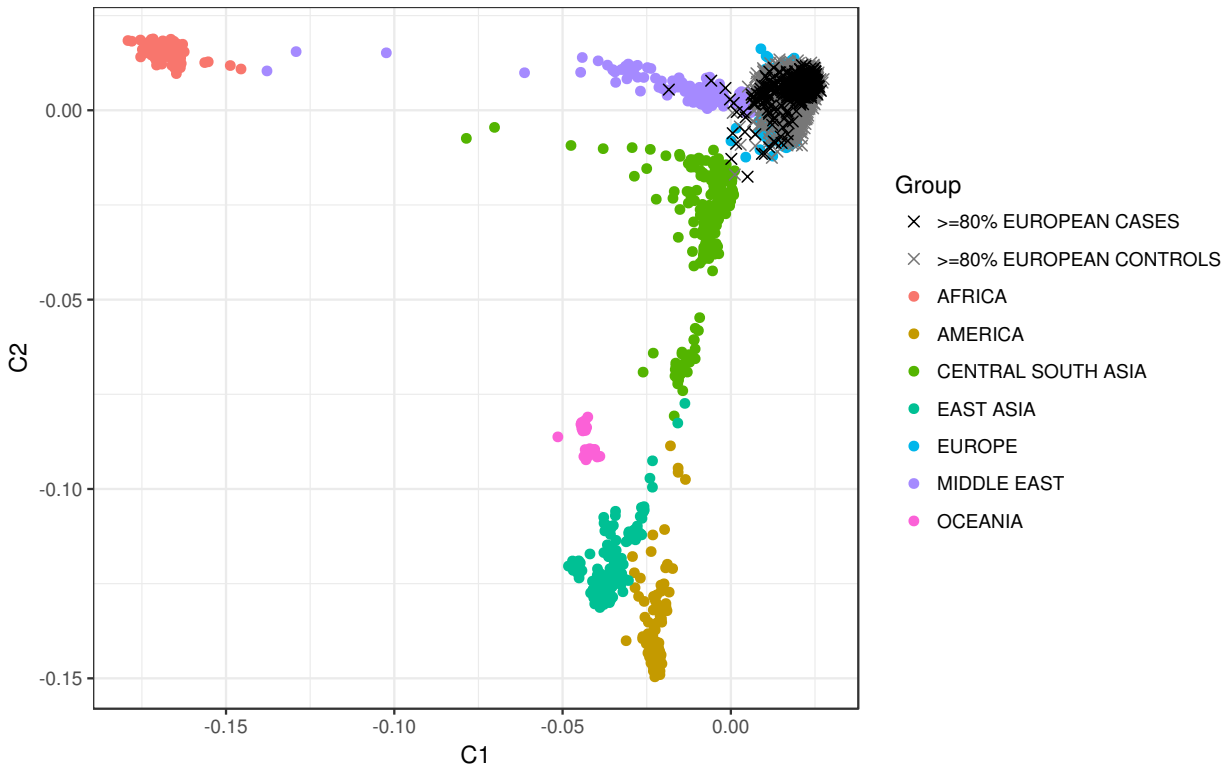
miRNA	Genes	Known miRNA expression associations in MS
hsa-miR-141	*CD80, THAP5*	
hsa-miR-197	*CD109, TSEN2*	Decreased in T cells of patients treated with IFN-β[6]
hsa-miR-200c	*SLC35B4*	hsa-miR-200c increased in white matter[7]; hsa-miR-200a and hsa-miR-200b decreased in B cells[7]; hsa-miR-200a decreased in whole blood[8]
hsa-miR-21	*C11orf70, PLAA*	Increased in white matter[7]; decreased in peripheral blood of ped-MS cases[14]
hsa-miR-3128	*CBL, SCLY*	
hsa-miR-3188	*PRSS12*	
hsa-miR-3605	*ARL6IP6*	Increased in peripheral blood of ped-MS cases[14]
hsa-miR-4277	*ZNF286B*	
hsa-miR-4294	*SLC37A4*	
hsa-miR-4498	*NCS1, RAB35*	
hsa-miR-4649	*HYOU1*	
hsa-miR-587	*PRKRIR, UTP18*	
hsa-miR-599	*PAPPA*	Increased in PBMCs[6] and decreased in B cells[7]
hsa-miR-608	*ADPRH, CD109, CIITA, COX10, CYB561D1, EHD2, GAST, GLIS2, HYOU1, NTSR1, PHF19, PIWIL3, PXN, SNAI1, SYNJ2BP, TFAP4, ZSCAN20*	
hsa-miR-744	*TANC2*	Increased in PBMCs[6]
hsa-miR-875	*EIF5A2, ERP29*	

Pairs are candidate biomarkers if both the miRNA and target gene are nominally associated with ped-MS ($p<0.01$) and the miRNA-target gene binding prediction score is in the top one percentile of all pairs. The last column indicates previously observed MS associations in miRNA expression studies.

Table 4: Pathways in which the 255 protein-coding genes containing miR-SNPs associated with ped-MS ($p<0.01$) are statistically overrepresented.

Pathway name	# Protein-coding genes in pathway	# Expected in 255 ped-MS genes	# Found in 255 ped-MS genes	P-value (FDR)	Ped-MS genes in pathway
Histamine H_1 Receptor mediated signaling	43	0.58	5	0.035	*GNG4, PLCB3, PLCG2, PRKCB, PRKCI*
5-HT_2 type receptor mediated signaling	66	0.86	6	0.064	*PLCB3, PRKCI, PLCG2, GNG4, PRKCB, SLC18A2*
MHC Protein Complex	25	0.34	5	0.085	*HLA-DPB1, HLA-DQB1, HLA-DRA, HLA-A, HLA-G*
Integral component of lumenal side of endoplasmic reticulum membrane	28	0.38	5	0.046	*HLA-DPB1, HLA-DQB1, HLA-DRA, HLA-A, HLA-G*
Interferon gamma signaling	90	1.21	9	0.014	*HLA-DPB1, TRIM14, HLA-DQB1, HLA-DRA, HLA-A, TRIM26/AFP, TRIM10, HLA-G, CIITA*

For each pathway, the total number of protein-coding genes in the pathway is given, followed by the number of those genes expected by chance to be found among the 255 ped-MS associated protein-coding genes, the actual number found, the p-value for statistical overrepresentation (adjusted for multiple hypothesis tests), and the list of ped-MS associated genes in the pathway.



Supplementary Figure: Multidimensional scaling plot of ped-MS cases (black) and controls (gray) with ≥80% European ancestry who were included in the GWAS, along with individuals from the Human Genome Diversity Project.

Chapter 5 - TNFa drives DNA methylation and transcriptional changes and activates autoimmune disease genes in endothelial cells

Abstract

Endothelial cells are a primary site of leukocyte recruitment during inflammation. An increase in tumor necrosis factor-alpha (TNFa) levels as a result of infection or as a result of some autoimmune diseases can trigger this process. Several autoimmune diseases are now treated with TNFa inhibitors. However, the genomic alterations that occur as a result of TNF-mediated inflammation are not well understood. To investigate the molecular targets and networks resulting from increased TNFa, we measured DNA methylation and gene expression in 40 human umbilical vein endothelial cell (HUVEC) primary cell lines before and 24 hours after stimulation with TNFa via microarray. Weighted gene co-expression network analysis (WGCNA) identified 15 groups of genes (modules) with similar expression correlation patterns, and four modules showed a strong association with TNFa treatment. Genes in the top TNFa-associated module were all up-regulated, had the highest proportion of hypomethylated regions, and were associated with 136 Disease Ontology terms, including autoimmune/inflammatory, infectious and cardiovascular diseases, and cancers. Another module was associated with cardiovascular and metabolic diseases but not TNFa, which is notable because cardiovascular diseases are increased in some autoimmune diseases. Of 223 hypomethylated regions identified, 28 were in gene promoters, and several of those genes have previously been associated with autoimmune disease in GWAS. These results reveal specific groups of genes that act in concert in endothelial cells and delineate those driven by TNFa and establish their relationship to DNA methylation changes, which has strong implications for understanding disease etiology and precision medicine approaches to disease therapy.

Significance Statement

TNFa is a cell-signaling protein involved in a wide variety of normal biological functions, including response to infections and inhibition of tumor formation. It is present at abnormally high levels in autoimmune diseases, and several autoimmune diseases are treated with TNFa inhibitors. However, these drugs do not work perfectly and can have unwanted side effects, so a better understanding of the effects of TNFa in various cell types is needed. This study characterized gene expression and DNA methylation changes in endothelial cells treated with TNFa. These cells line the interior surface of blood vessels and lymphatic vessels and are at the interface between the blood and autoimmune disease targets. These results can guide future research into improving autoimmune disease therapy.

Introduction

TNFα is an inflammatory cytokine that is dysregulated in many autoimmune diseases and is generally found at increased levels in disease-relevant tissues. TNFa inhibitors are a

major class of treatment for autoimmune diseases, including rheumatoid arthritis, psoriasis, psoriatic arthritis, inflammatory bowel disease, ulcerative colitis, Crohn's disease, and ankylosing spondylitis.(1) However, TNFa inhibitors are imperfect treatments with several side effects, such as increased risk of infections and non-melanoma skin cancers, and they can even induce autoimmune disease, including lupus, psoriasis, and CNS demyelination.(2–6) TNFa has a range of biological functions that can be either homeostatic, e.g., defense against pathogens, tissue regeneration, immunoregulation, and inhibition of tumor formation, or pathogenic, e.g., recruitment of inflammatory cells, inhibition of T regulatory cells, necroptosis, and tissue degeneration.(7) Because so many (sometimes contradictory) biological processes are activated by TNFa and disrupted by TNFa inhibitors, there is a need to move from therapies that globally influence TNFa toward therapies that can pinpoint its pathogenic processes while leaving homeostatic processes undisturbed.

TNFa plays different roles in different cell types. Endothelial cells, found in the linings of blood vessels, are of particular interest in autoimmune disease because they directly interact with leukocytes to bring them to sites of inflammation or infection.(8) Endothelial cells are also of interest because dysfunction of these cells is more common in those with autoimmune disease, causing accelerated atherosclerosis and other cardiovascular disease, making it a leading cause of mortality among patients, especially in rheumatic autoimmune diseases.(9–11)

In this study, DNA methylation and gene expression changes were characterized in human umbilical vein endothelial cell (HUVEC) primary cell lines after treatment with TNFα in order to help identify new therapeutic targets and to provide information that can be used to help predict possible side effects. A systems biology approach, weighted gene co-expression network analysis (WGCNA), was used to construct a gene expression network and find groups of genes that not only have a high correlation of expression but high topological overlap, meaning that to be considered members of the same group, genes need to show similar correlation patterns to other genes outside of the group. Each group, or module, was then tested to determine whether expression was associated with TNFa treatment. This strategy greatly reduces the multiple hypothesis testing burden, allows the identification groups of genes that act in a coordinated fashion, and reveals groups of transcripts that are differentially expressed in response to TNFa stimulation.

To further understand how TNFa affects endothelial cells, differentially methylated regions (DMRs) were identified. DNA methylation affects gene transcription in different ways depending on where it is located, though the relationship between methylation and expression is still not entirely understood. Increased methylation in promoter regions is the most well studied and generally induces stable repression of gene expression, while increased methylation in gene bodies frequently coincides with increased expression. Decreased methylation in enhancers is mostly associated with increased transcription factor binding.(12, 13) DMRs were related to genes in WGCNA modules by identifying DMRs within genes. GeneHancer (GH), a database of promoters and enhancers and their inferred target genes(14) was used to reveal DMRs in promoter and enhancer regions.

The transcription factor NF-kappa-B (NF-κB) was of particular interest in this study because its activation is one of the major mechanisms by which TNFa exerts its effects. TNFa increases NF-κB expression, which in turn regulates a host of genes involved in inflammation and immune responses.(15) The GH elements for genes in WGCNA modules were overlaid with known HUVEC-specific NF-κB transcription factor binding sites (TFBSs) to identify genes that are likely to be regulated by this transcription factor in endothelial cells and to gauge whether NF-κB is a master regulator of specific modules.

Finally, Disease Ontology enrichment analysis was performed on genes in each WGCNA module to elucidate the known associations of genes to disease.

Results

Differential expression and identification of gene modules

WGCNA based on a signed network identified 15 gene modules with high topological overlap; i.e., 15 clusters of genes with similar patterns of connection to other genes (Figure 1). Each module is designated by a color, and the expression pattern of each module is summarized by the "module eigengene," which is the first principal component of expression for all genes in the module. The relationship of module eigengenes to one another is shown in Figure 2. This relationship shows that, for example, expression of genes in the green module is positively correlated with those in the black module, but uncorrelated with those in the cyan module. The number of genes per module ranged from 34 to 2,570, and roughly 10% of genes were not part of any module, but collected in the grey "module" (Table 1). To understand the effect of TNFa on genes in each module, two approaches were used: (1) genes that were significantly up- or down-regulated according to moderated paired *t*-tests were identified for each module, and (2) each of the 15 module eigengenes (MEs) was tested for association with TNFa treatment in linear mixed regression models. Of 14,019 genes detected in HUVEC cell lines, 3,060 were upregulated with TNFa and 5,089 were downregulated (Supplementary File 1). The green, purple, black, and brown modules were highly associated with TNFa treatment, with Bonferroni-adjusted $p < 10^{-15}$. Genes in the green and black modules nearly all showed increased expression with TNFa, while the purple and brown genes showed decreased expression. Six modules, turquoise, greenyellow, tan, red, salmon, and yellow, were moderately negatively associated with TNFa treatment ($0.05 < \text{adjusted } p > 10^{-15}$) and contained more down-regulated than up-regulated genes. The remaining five modules were unassociated with TNFa. Reassuringly, the grey module, which contains the collection of unassigned genes, was not associated with TNFa treatment. In all cases, the sign of the ME regression coefficient corresponded with whether the majority of genes was up- or down-regulated (i.e., a positive regression coefficient corresponded with majority up-regulated genes). The full list of genes and module assignments is available in Supplementary File 1.

Comparison of module genes to methylation changes and TFBSs

Bumphunter identified 223 differentially methylated regions associated with TNFa treatment, all hypomethylated (Supplementary File 2). Of these, 131 DMRs were located within a gene, 186 were located in GeneHancer (GH) regulatory elements (categories are not mutually exclusive; 109 DMRs were in both), and 28 were located specifically in GH promoters (Supplementary File 2). Of the genes with DMRs in their promoters, 17 also contained SNPs (131 unique) associated in GWAS with several traits, including autoimmune, cardiovascular, and metabolic diseases. The relationship of WGCNA gene modules and DMRs is shown in Table 2. The most highly TNFa-associated module, green, contained 34 genes (3.2%) with DMRs, which was more than any other module. The green module also had one of the highest proportions (0.29%) of DMRs in gene-related GH regulatory elements; cyan and midnightblue also had high proportions, but of a much smaller number of GH elements. In general, more DMRs were present in GH enhancers than in GH promoters.

For most modules, about 10% of the mapped GH elements overlapped binding sites for the NF-κB p65 subunit encoded by RELA (range: 7.7%-11.1%). The proportion of elements with both a DMR and RELA TFBS was slightly less than the proportion of elements with only a DMR, meaning that most GeneHancer elements with a DMR are known RELA TFBSs (Supplementary Table 1).

Disease Ontology of module genes

Module genes were overrepresented among genes assigned to Disease Ontology (DO) terms for the green, black, and cyan modules. A total of 136 DO terms were enriched for genes in the green module (up-reguated with TNFa), with infectious, respiratory, skin, connective tissue, and hypersensitivity reaction diseases being among the most statistically significant, but several autoimmune diseases, including systemic lupus erythematosus, rheumatoid arthritis, Graves' disease, psoriasis, and multiple sclerosis were also enriched for green module genes, as were diabetes, coronary artery disease, and atherosclerosis (Figure 3, Supplementary Figure 1, and Supplementary File 3). Lupus genes were overrepresented in the black module (up-regulated with TNFa), along with skin and pleural cancers, nephritis, and purpura (Supplementary Figure 2 and Supplementary File 3). Cardiovascular and metabolic diseases, including coronary artery disease, atherosclerosis, diabetes, and obesity, were overrepresented in the cyan module, which was not significantly associated with TNFa treatment (Supplementary Figure 4 and Supplementary File 3).

Discussion

In this experimental study, sets of genes related by similar expression patterns in endothelial cells were identified, and the extent to which expression changed as a result of TNFa stimulation was estimated. Three sets of genes identified by WGCNA were overrepresented among established Disease Ontology genes, and two of those sets, green and black, were associated with response to TNFa stimulation. Specific genes

associated with each disease may be used to help explain the mechanisms by which global changes to TNFa levels affect many phenotypes and risk for multiple diseases. Autoimmune, cardiovascular, metabolic, and cancer disease processes occurring in endothelial cells as a result of increased TNFa are likely to be driven by the genes in the green and black modules. Interestingly, the cyan module genes, which were associated with obesity, diabetes, coronary artery disease, and atherosclerosis, were not associated with TNFa, suggesting that while the genes in the cyan list are acting together, they are probably not being driven by TNFa changes. This is the first study to characterize the gene expression network of TNFa-stimulated endothelial cells.

Supplementary File 1 contains the full lists of genes in each module, along with measures of how connected each gene is to other genes. These lists can be used to help predict, along with other resources such as the STRING database,(16) whether targeting specific genes therapeutically is likely to have an effect on many other genes or not. For instance, the top 10 most highly connected genes within the green module are *TAP1, CX3CL1, CXCL10, PSME2, EBI3, UBD, TNFAIP3, PSMB9, SLC15A3,* and *TNFRSF9*. Because expression of these "hub genes" is highly correlated with many other genes, they are likely to be integral to the regulation of those other genes, while genes with low connectivity are less likely to be tightly coupled to many other genes. There are several gene measures reported in Supplementary File 1, each with slightly different meanings and implications, though the gene measures tend to be highly correlated.(17) Connectivity refers to the sum of connection (correlation) strengths with other genes in the network. The measure *kWithin* is the intramodular connectivity, or connectivity of a particular gene to all other genes within its same module and *kTotal* is connectivity to all other genes regardless of module (*kOut* is kTotal-kWithin, and *kDiff* is kWithin-kOut). A gene with a high kWithin measure but a low kTotal measure is one that is connected mainly to genes only within its module and could therefore reasonably be expected not to affect genes in other modules. Module membership (MM) is the correlation of a gene's expression to the module eigengene (the first principal component of expression level of all genes in the module) and is an indicator of how representative expression of that gene is to the other genes in the module. MM can be calculated for both the module a gene belongs to and all other modules. Gene significance (GS) is the association of a gene's expression level with treatment with TNFa, and it is the only measurement that is directly tied to TNFa. Ideally, candidate genes for future therapeutic research would have a high GS measure and a low kTotal measure, indicating that the genes are affected by TNFa fluctuations but that they are not likely to affect many other genes.

Genes in the green module are of particular interest for further study because they were the most highly associated with response to TNFa stimulation, both individually (high GS values), and as a group (strongest and most significant association of the module eigengene with TNFa treatment), and genes in this module were by far the most overrepresented for diseases in the Disease Ontology database. These included genes such as chemokines *CXCL1*, *CXCL10* and *CXCL8* and genes associated with autoimmune diseases such as *HLA-C*, *DDX58*, *IL4*, *NFKBIA* and *TNFAIP3* which are associated with psoriasis susceptibility. Moreover, NFKB1 from this module mediates

Th1/Th17 activation in the pathogenesis of psoriasis and probably other autoimmune diseases.(18) Th17 activation is particularly significant for the development of a number of autoimmune diseases. Green genes also had more DMRs, either in the gene bodies or in GH elements mapped to the genes, suggesting that TNFa stimulation causes more long-lasting changes to gene expression to genes in the green set than to other sets. It should be pointed out that the WGCNA results were based on a signed network, which treats strongly negatively correlated genes as unconnected and means that genes in each group are generally positively correlated with one another. This also means that the genes in the top TNFa-associated modules are almost all up-regulated with TNFa (e.g., green, black) or down-regulated with TNFa (e.g., purple, brown). Genes that are strongly negatively correlated, for instance, genes that inhibit the expression of other genes, are not captured in the same module. The relationship of genes in different modules is captured by the MM measures for all genes in all modules (see Supplementary File 1). These MM measures for genes in different modules can be used to understand which genes in, e.g., the green module are strongly negatively correlated with most genes in, e.g., the purple module. Sets of genes in both the green and cyan modules were overrepresented in cardiovascular and metabolic diseases. These sets may be useful in future studies that aim to explain the overlap of obesity, autoimmune disease and cardiovascular disease.(19–21) The green module genes were overrepresented in all three types of disease, were strongly associated as a set with TNFa, and were nearly all up-regulated by TNFa, while the cyan module genes were overrepresented in metabolic and cardiovascular disease but not autoimmune disease, were not associated as a set with TNFa, and were mostly not up-regulated by TNFa. It is therefore possible that it is the green module genes, and not the cyan module genes, that are driving the overlap among these disease types, but further investigation is needed.

NF-κB binding sites were generally evenly distributed among the enhancers for genes all of the WGCNA gene sets, suggesting that NF-κB is not a master regulator of any specific modules. It should be noted that NF-κB consists of a collection of different heterodimers of seven proteins, but RELA/p65 is considered the prototypical form.(22) It is possible that if binding of the other proteins were measured by ChIP-Seq, a slightly different picture of NF-κB in endothelial cells would emerge.

This study had some limitations. In particular, several steps required relating data types to one another based on gene symbols (common names), which is an imperfect process, as genes may have multiple names and change over time. In addition, not all diseases and disease-gene associations are captured by the Disease and Gene Annotations database, so some diseases may have been missed in the Disease Ontology enrichment analysis. Most of the methods employed, especially WGCNA, require selecting specific settings that, when adjusted, may change the final results somewhat. Finally, while this was a well-powered study in primary endothelial cells*, in vivo* results may be somewhat different, and other cell types would need to be evaluated to get a more complete picture of how these genes are affected by TNFa.

In summary, this study utilized a comprehensive systems biology approach integrating multiple data types and state of the art bioinformatics tools to reveal groups of correlated genes with similar patterns of expression in endothelial cells. One of these groups of genes is highly associated with TNFa and with cancers and infectious, autoimmune and cardiovascular diseases. Another group is not responsive to TNFa but plays an important role in metabolic and cardiovascular diseases. The detailed results provided in supplementary files can inform future research on new drug targets for diseases that are currently treated with TNFa inhibitors.

Materials and Methods

Samples and Data Generation

Forty primary human umbilical vein endothelial cell (HUVEC) lines were obtained from Promocell and were cultured until passage four. Each cell line was split in half, and one half was treated with 20ng/mL TNFa for 24 hours while the other half was left untreated. Cells were then pelleted, and DNA and RNA were isolated. Gene expression was measured with Illumina HT-12 V4 expression BeadChip microarrays by Eurofins Genomics, and DNA methylation was measured with Illumina Infinium MethylationEPIC BeadChip microarrays.

Gene Expression Analysis

Gene expression data were processed with the *limma* R package.(23) Outlier samples were detected with boxplots and classical multidimensional scaling (MDS) plots of the log2 probe intensities and removed from analysis. The *neqc* function was used to perform background correction and quantile normalization, log2 transformation of the probe intensities, and removal of control probes. Probes with expression detected (detection $p<0.05$) in less than half of the samples (n=28,386) were removed from analysis. The remaining 18,937 probes were collapsed to 14,019 genes by selecting the probe with the maximum mean intensity value for each gene using the *collapseRows* function. After quality control, 39 sample pairs (treated and untreated) remained for analysis. Differential expression was determined with linear regression models with log2 intensity as the outcome and treatment with TNFa and HUVEC pair identifier as the predictors. A moderated paired *t*-test statistic for each gene was computed with the empirical Bayes method in *limma*.

Weighted gene correlation network analysis (WGCNA)

The *WGCNA* R package was used to construct a correlation network of genes, identify gene modules consisting of interconnected genes, study module relationships, and find the key drivers of each module.(17) A signed co-expression network was constructed from all 14,019 genes expressed in HUVEC cells, using data from all 78 treated and untreated samples together. A soft-thresholding power of $\beta=16$, a minimum module size of 30, and a merge cut height of 0.25 were used. Eigengenes (the first principal

component of gene expression values) were determined for each gene module. Because of the paired design of the study, linear mixed-effect models were used to estimate the relationship of module eigengenes to TNFa treatment.(24) Linear mixed effect models were used to estimate the relationship of each module eigengene to TNFa treatment with the *lmer* function in the *lme4* R package(25), considering treatment as a fixed effect and HUVEC pair identifier as a random effect. Gene significance, the association of each gene with TNFa treatment, was determined with the same linear mixed effect models described above, using expression of individual genes as the outcomes instead of the module eigengenes. Hub genes were selected based on connectivity calculated with *WGCNA*, and module membership (the correlation of each gene with its module eigengene) and gene significance were also reported.

DNA methylation analysis

The *minfi* R package was used for data preprocessing, normalization, and quality control of DNA methylation data.(26) Background subtraction and dye bias correction was performed using the *preprocessNoob* function, followed by quantile normalization with *preprocessQuantile*. Samples with more than 5% poor detection p-values (>0.01) were removed from analysis. CpG sites with poor detection p-values across samples were removed from analysis (n=3,451 sites). Predicted sex based on X and Y chromosome methylation was checked against recorded sex. CpG sites with probes predicted to cross-hybridize to other genomic locations were removed from analysis (n=44,032 sites).(27) The final dataset used for analysis consisted of 37 sample pairs and 818,391 CpG sites. Differentially methylated regions (DMRs) were identified using the *bumphunter* R package.(28) Bumphunter was run using the same linear regression models as the expression analysis, using methylation M-values as the outcome and TNFa and HUVEC pair identifiers as predictors. CpG sites were considered to be part of a cluster if they had no more than 1,000 bases between them. One thousand bootstrap samples were used to generate a null distribution of regions. Candidate differentially methylated regions were nominated with pickCutoff, using the 99% quantile of the null-distribution as a threshold value, and a family-wise error rate (FWER) cutoff of 0.05 was used to determine statistical significance. DMRs were mapped to RefSeq Genes by intersecting DMR positions with genes in the *ncbiRefSeq* table using the UCSC Genome Browser (GRCh37/hg19 assembly coordinates).(29, 30)

Comparison of microarray results and mapping to GeneHancer elements and TFBSs

In order to compare gene expression results to methylation results, and to facilitate annotation, each probe from the Illumina HT-12 V4 microarray was assigned a current gene name using microarray probe mappings from Ensembl (Genebuild v96).(31) Of the 47,231 probes on the microarray, 37,525 (79%) mapped to a stable Ensembl Gene identifier. After removing probes that mapped to multiple genes and collapsing the coordinates of genes with multiple transcripts by taking the minimum transcription start coordinate and maximum transcription end coordinate, 34,094 probes remained, corresponding to 22,628 genes.

In addition to obtaining up-to-date gene names for the Illumina microarray, promoter and enhancer elements were identified for most genes using the GeneHancer track on the UCSC Genome Browser (GRCh37/hg19 assembly).(14) Genes were matched to GeneHancer elements using one of: (1) the Ensembl v96 gene symbol obtained from the Illumina probe mapping above, (2) the Ensembl v92 gene symbol (as the Ensembl v92 regulatory build was used to generate the GeneHancer track), or (3) the Ensembl v92 stable gene identifier, for GeneHancer elements with no gene symbol. In total, 31,625 Illumina expression probes and 20,712 genes were successfully mapped to GeneHancer elements.

Finally, RELA TFBSs in TNFa-treated HUVEC cell lines were compared to both methylation and expression results. TFBSs from three Chip-Seq datasets deposited in GEO (GSE53998, GSE34500, and GSE43070) and uniformly processed using ChiP-eat software and the PWM peak caller were downloaded from the UniBind website, https://unibind.uio.no/.(32)

DMRs within genes were compared to genes in WGCNA modules using gene symbols. DMRs within GH elements and RELA TFBSs were identified by intersecting their positions using BEDTools.(33) DMRs in GH promoters were identified using the UCSC Genome Browser, and GWAS information for genes with the closest (usually overlapping) 5'UTRs to the GH promoters was extracted using the UCSC Data Integrator tool and the GWAS Catalog track.(30, 34)

Disease Ontology enrichment analysis

Disease Ontology enrichment analysis was performed on genes in each module using the *XGR* R package,(35) which utilizes the Disease and Gene Annotations database to map genes to diseases.(36) Gene symbols from the Ensembl Genebuild (v96) were used as input to XGR when available; otherwise, original gene symbols from the Illumina HT-12 V4 expression microarray manifest file were used. The list of 14,019 genes expressed in HUVEC cells was supplied as the background gene list for enrichment tests. To be considered as an enriched disease term, at least 10 and at most 2,000 genes were required to be annotated for that term, and at least 5 genes were required to overlap with the input gene list. Fisher's exact test was used to determine significance, and parent-child relations were accounted for using the "lea" algorithm. Disease terms enriched at an FDR-adjusted $p<0.05$ were reported.

Acknowledgments

This work was supported by NIH grants AR050266 to A.M.B, R01ES017080 to L.F.B, and F31NS096885 to B.R.

References

1. Li P, Zheng Y, Chen X (2017) Drugs for autoimmune inflammatory diseases: From small molecule compounds to anti-TNF biologics. *Front Pharmacol* 8(JUL):1–12.
2. Murdaca G, et al. (2015) Infection risk associated with anti-TNF-α agents: a review. *Expert Opin Drug Saf* 14(4):571–582.
3. Lindhaus C, Tittelbach J, Elsner P (2017) Cutaneous side effects of TNF-alpha inhibitors. *JDDG - J Ger Soc Dermatology* 15(3):281–288.
4. Williams EL, Gadola S, Edwards CJ (2009) Anti-TNF-induced lupus. *Rheumatology* 48(7):716–720.
5. Brown G, et al. (2017) Tumor necrosis factor-α inhibitor-induced psoriasis: Systematic review of clinical features, histopathological findings, and management experience. *J Am Acad Dermatol* 76(2):334–341.
6. Kemanetzoglou E, Andreadou E (2017) CNS Demyelination with TNF-α Blockers. *Curr Neurol Neurosci Rep* 17(4). doi:10.1007/s11910-017-0742-1.
7. Kalliolias GD, Ivashkiv LB (2016) TNF biology, pathogenic mechanisms and emerging therapeutic strategies. *Nat Rev Rheumatol* 12(1):49–62.
8. Langer HF, Chavakis T (2009) Leukocyte - Endothelial interactions in inflammation. *J Cell Mol Med* 13(7):1211–1220.
9. Ahearn J, Shields KJ, Liu CC, Manzi S (2015) Cardiovascular disease biomarkers across autoimmune diseases. *Clin Immunol* 161(1):59–63.
10. Atehortúa L, Rojas M, Vásquez GM, Castaño D (2017) Endothelial Alterations in Systemic Lupus Erythematosus and Rheumatoid Arthritis: Potential Effect of Monocyte Interaction. *Mediators Inflamm* 2017(Ic):1–12.
11. Durante A, Bronzato S (2015) The Increased Cardiovascular Risk in Patients Affected by Autoimmune Diseases: Review of the Various Manifestations. *J Clin Med Res* 7(6):379–384.
12. Luo C, Hajkova P, Ecker JR (2018) Dynamic DNA methylation: In the right place at the right time. *Science (80-)* 361(6409):1336–1340.
13. Tirado-Magallanes R, Rebbani K, Lim R, Pradhan S, Benoukraf T (2017) Whole genome DNA methylation: beyond genes silencing. *Oncotarget* 8(3):5629–5637.
14. Fishilevich S, et al. (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* 2017:1–17.
15. Zhang Q, Lenardo MJ, Baltimore D (2017) 30 Years of NF-κB: A Blossoming of Relevance to Human Pathobiology. *Cell* 168(1–2):37–57.
16. Szklarczyk D, et al. (2017) The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 45(D1):D362–D368.
17. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9(559). doi:10.1186/1471-2105-9-559.
18. Zhou F, et al. (2018) NFKB1 mediates Th1/Th17 activation in the pathogenesis of psoriasis. *Cell Immunol* 331(January):16–21.
19. Endo Y, Yokote K, Nakayama T (2017) The obesity-related pathology and Th17 cells. *Cell Mol Life Sci* 74(7):1231–1245.
20. Van Raemdonck K, Umar S, Szekanecz Z, Zomorrodi RK, Shahrara S (2018)

Impact of obesity on autoimmune arthritis and its cardiovascular complications. *Autoimmun Rev* 17(8):821–835.
21. Granata M, et al. (2017) Obesity, Type 1 Diabetes, and Psoriasis: An Autoimmune Triple Flip. *Pathobiology* 84(2):71–79.
22. Chen LF, Greene WC (2004) Shaping the nuclear action of NF-κB. *Nat Rev Mol Cell Biol* 5(5):392–401.
23. Ritchie ME, et al. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43(7):e47.
24. Li J, et al. (2018) Application of Weighted Gene Co-expression Network Analysis for Data from Paired Design. *Sci Rep* 8(1):1–8.
25. Bates D, Mächler M, Bolker BM, Walker SC (2015) Fitting Linear Mixed-Effects Models using lme4. *J Stat Softw* 67(1):1–48.
26. Aryee MJ, et al. (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30(10):1363–9.
27. McCartney DL, et al. (2016) Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genomics Data* 9:22–24.
28. Jaffe AE, et al. (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol* 41(1):200–209.
29. O'Leary NA, et al. (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44(D1):D733–D745.
30. Haeussler M, et al. (2019) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* 47(D1):D853–D858.
31. Cunningham F, et al. (2019) Ensembl 2019. *Nucleic Acids Res* 47(D1):D745–D751.
32. Gheorghe M, et al. (2019) A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Res* 47(4):e21.
33. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.
34. Buniello A, et al. (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 47(D1):D1005–D1012.
35. Fang H, Knezevic B, Burnham KL, Knight JC (2016) XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. *Genome Med* 8(1):1–20.
36. Peng K, et al. (2013) The disease and gene annotations (DGA): An annotation resource for human disease. *Nucleic Acids Res* 41(D1):553–560.

Tables and Figures

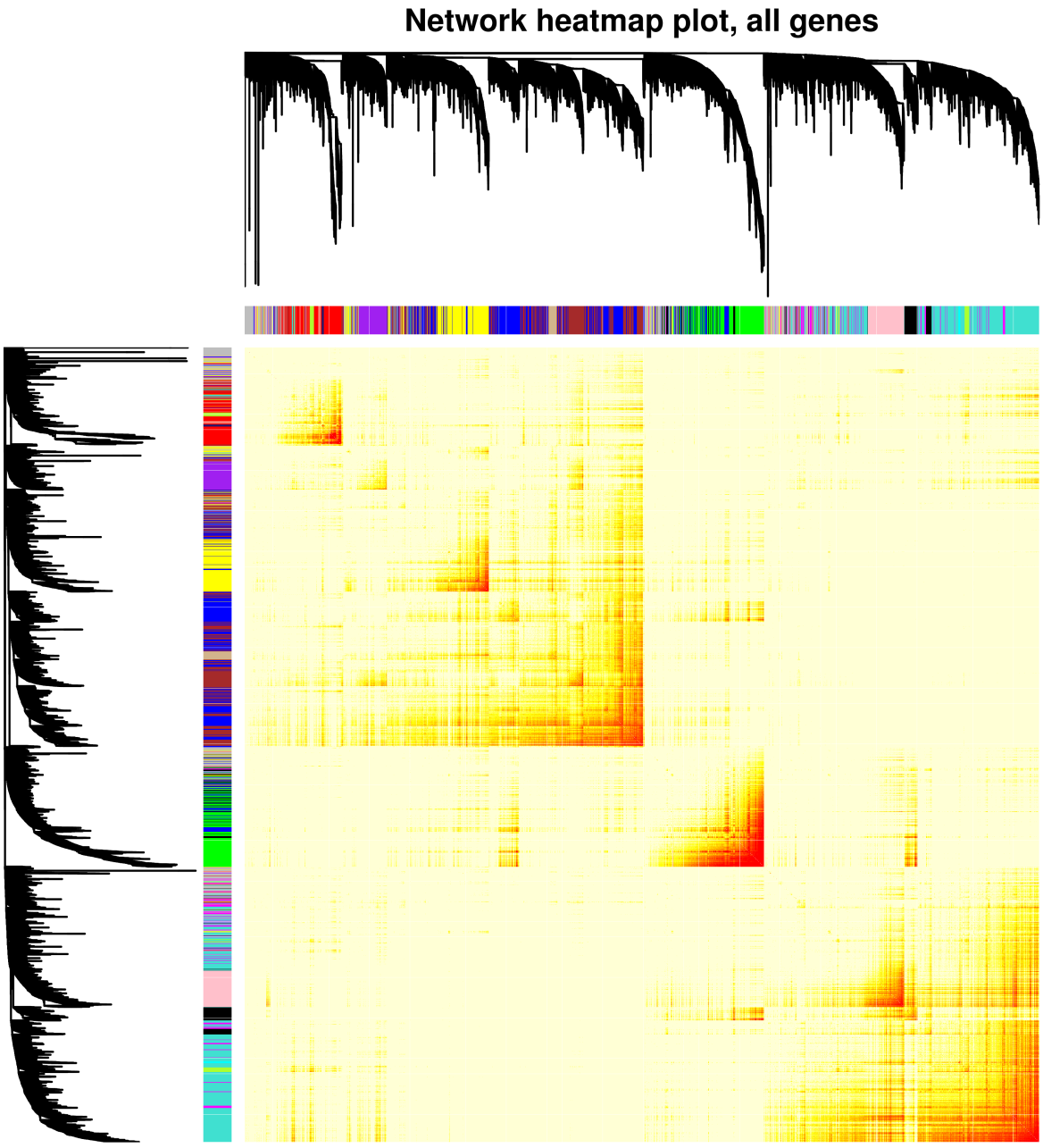


Figure 1. Visualization of the gene co-expression network modules. The co-expression network was built by considering all TNFa stimulated and unstimulated samples together. Hierarchical clustering dendogram of 14,019 genes expressed in HUVEC cell lines, along with colors representing module assignments. Genes that are not assigned to any module are colored grey. The heatmap shows the topological overlap matrix, and darker coloring indicates higher topological overlap.

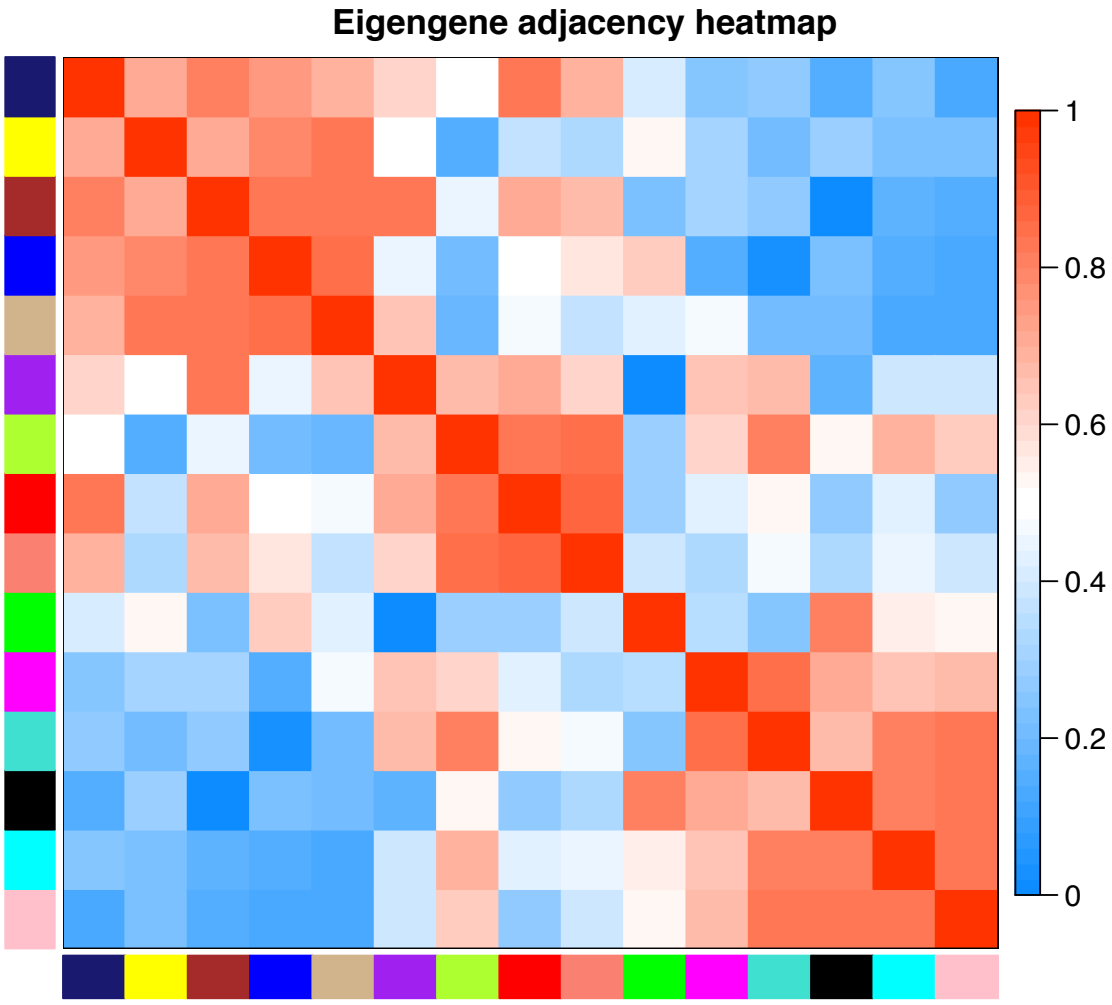


Figure 2. Visualization of the gene co-expression module relationships. Adjacency heatmap shows the relationships among the module eigengenes, which can be thought of as the weighted average gene expression of all genes in a module. For each pair of eigengenes E_I, E_J, adjacency is calculated as $(1 + \text{cor}(E_I, E_J))/2$. Red represents positively correlated modules and blue represents negatively correlated modules.

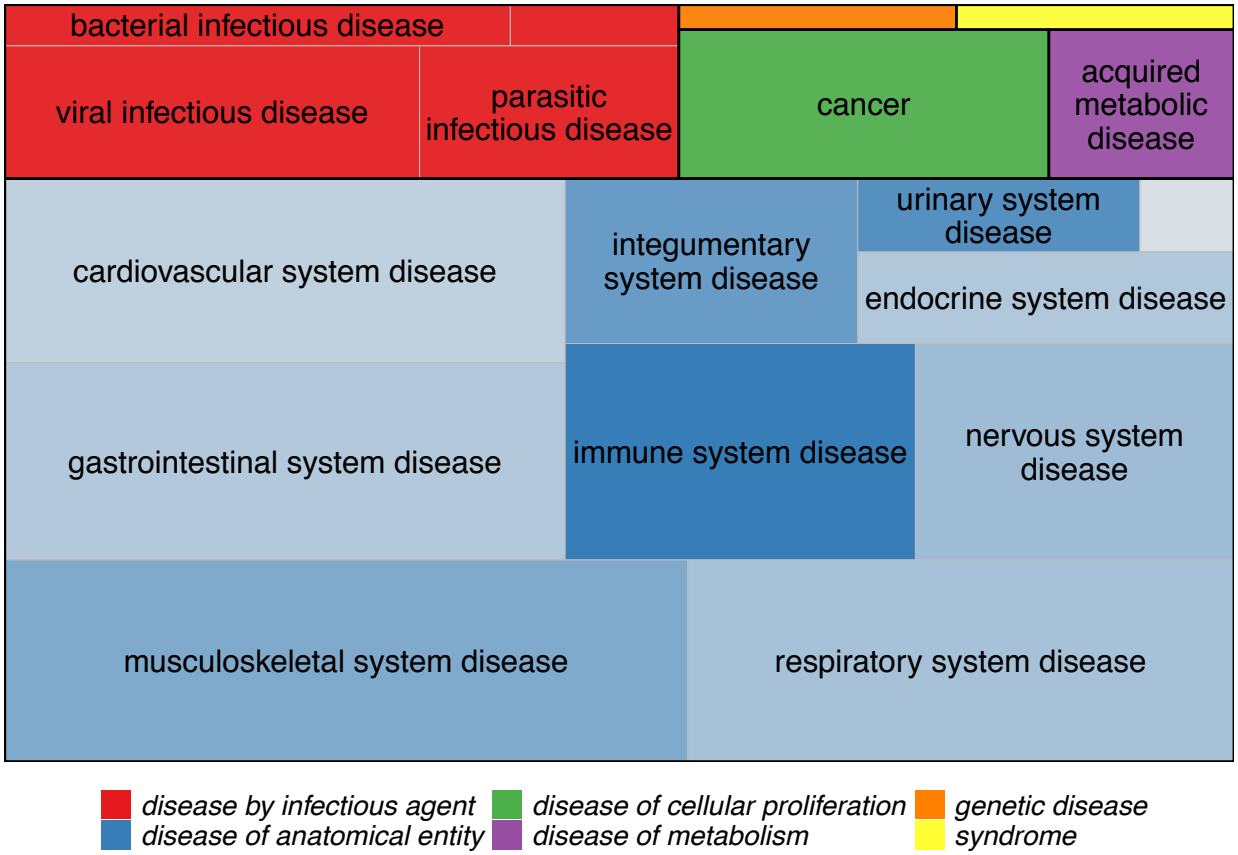


Figure 3. Disease Ontology categories for the 136 diseases that showed over-representation of green module genes. Boxes are colored according to the top-level disease categories and labels show the second-level categories. The size of each box is proportional to the number of disease terms in that category with significant overrepresentation of green module genes. Some disease terms belong to more than one category (e.g., multiple sclerosis is both a “nervous system disease” and an “immune system disease”), but each term is only represented once. For “disease of anatomical entity” terms, squares are shaded by the proportion of terms that represent autoimmune/inflammatory diseases (e.g., 3 of 16 “gastrointestinal system disease“ terms are autoimmune/inflammatory, while 9 of 11 “immune system disease” terms are). The full list of disease terms, with (manually curated) autoimmune/inflammatory terms highlighted, is given in Supplementary File 3.

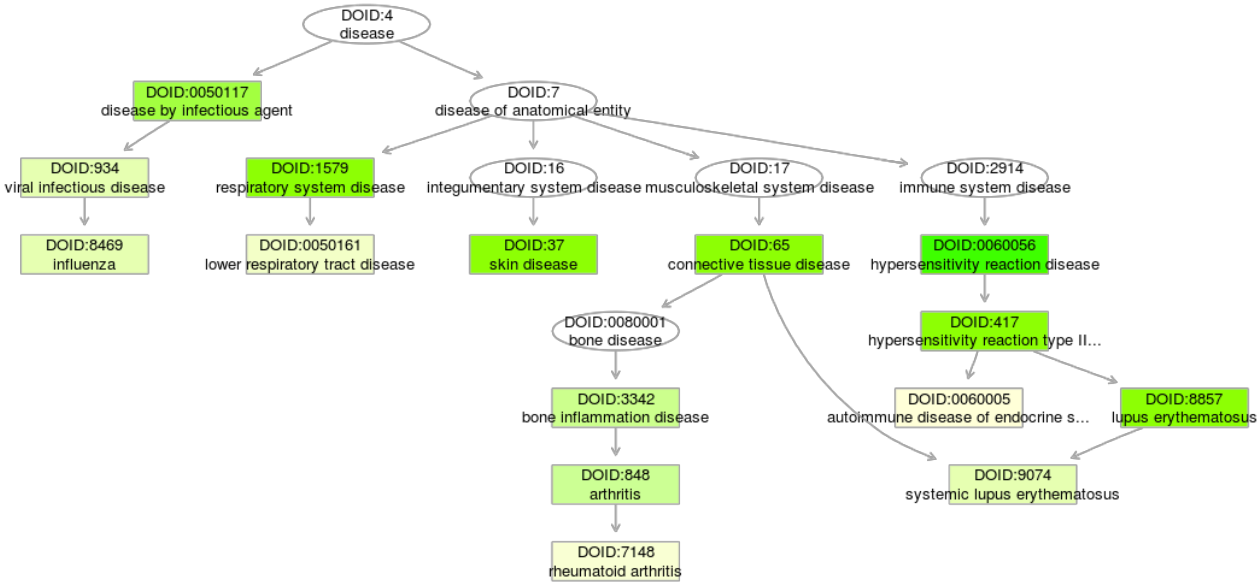
Module	Number of Genes	Up-regulated with TNFa	Down-regulated with TNFa	Beta value for TNFa association	P-value for TNFa association
Green	1,067	1,063	0	0.207	1.67E-31
Purple	491	0	490	-0.202	6.30E-28
Black	679	655	0	0.162	8.13E-20
Brown	1,633	0	1,487	-0.146	1.06E-15
Turquoise	2,570	132	1,203	-0.059	2.14E-07
Greenyellow	221	11	121	-0.054	1.76E-06
Tan	187	1	104	-0.062	7.66E-04
Red	828	29	431	-0.072	3.77E-03
Salmon	122	8	45	-0.037	5.78E-03
Yellow	1,252	100	386	-0.028	0.015
Cyan	37	16	0	0.046	0.090
Midnightblue	34	0	16	-0.052	0.119
Pink	673	263	14	0.042	0.339
Magenta	587	54	152	-0.026	0.363
Blue	2,202	587	421	0.013	1
Grey/unassigned	1,436	141	220	-0.005	1

Table 1. Description of gene modules identified by WGCNA and the relationship of genes in modules with TNFa. Each gene module is assigned a color, with grey reserved for genes not assigned to any module. The number of genes in each module is given, followed by the number of genes in that module that were significantly up- or down-regulated with TNFa according to individual gene tests in *limma*. The last two columns contain regression beta values and Bonferroni-adjusted p-values from linear mixed models testing the association of module eigengenes (the first principal component of gene expression in each module) and treatment with TNFa.

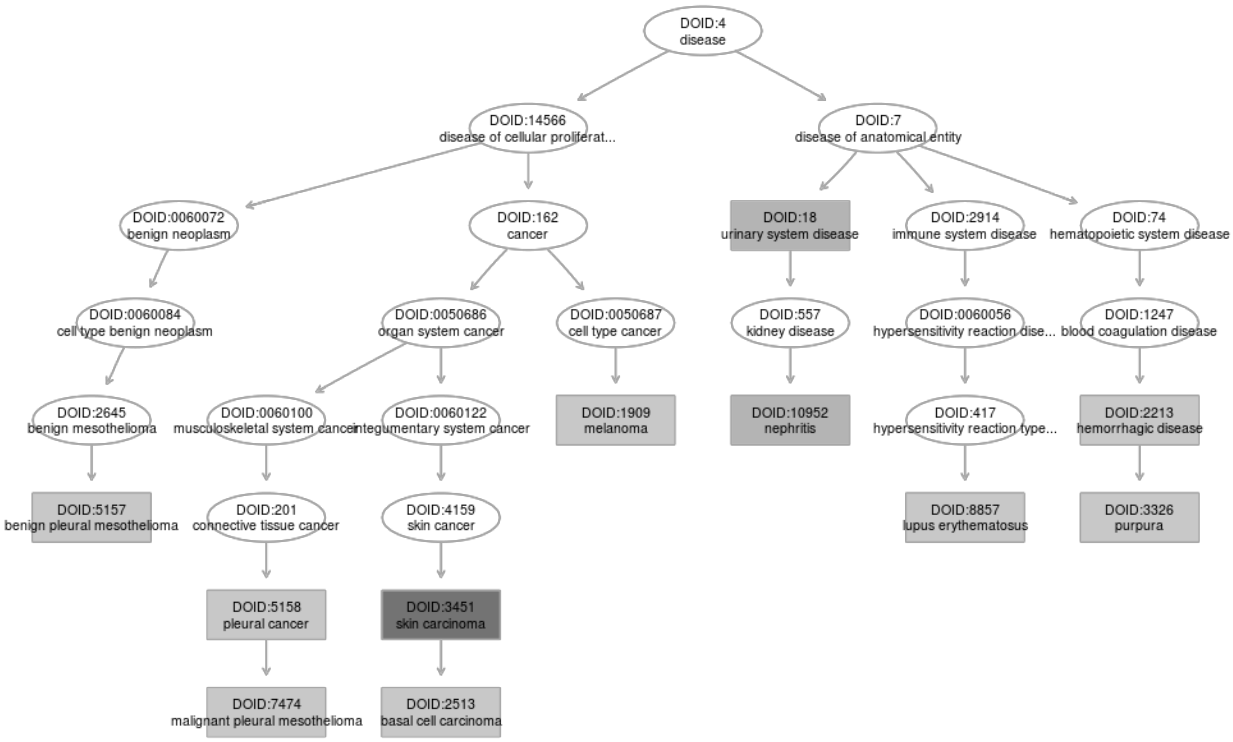
Module	Genes with DMRs	Percent of module genes	Total GH elements	GH elements with DMRs (promoter, enhancer)	Percent of module GH elements
Green	34	3.2%	29,456	85 (14, 71)	0.29%
Purple	0	0.0%	13,144	9 (0, 9)	0.07%
Black	7	1.0%	18,924	29 (2, 27)	0.15%
Brown	8	0.5%	44,637	43 (4, 39)	0.10%
Turquoise	9	0.4%	55,670	63 (9, 54)	0.11%
Greenyellow	3	1.4%	5,538	3 (0, 3)	0.05%
Tan	0	0.0%	3,918	2 (0, 2)	0.05%
Red	7	0.8%	20,300	20 (4,16)	0.10%
Salmon	2	1.6%	3,581	3 (1, 2)	0.08%
Yellow	5	0.4%	18,503	23 (7, 16)	0.12%
Cyan	0	0.0%	672	2 (0, 2)	0.30%
Midnightblue	0	0.0%	396	1 (0,1)	0.25%
Pink	0	0.1%	12,661	19 (3, 16)	0.15%
Magenta	2	0.3%	13,906	21 (7, 14)	0.15%
Blue	13	0.6%	59,406	68 (14, 54)	0.11%
Grey/unassigned	4	0.3%	30,967	28 (8, 20)	0.09%

Table 2. DMRs and their relationship to genes and GH elements in WGCNA modules. The number of genes overlapping a DMR in each module is given, followed by the percent of genes containing a DMR among all genes in the module, the total number of GH elements mapped to genes in the module, the number of GH elements overlapping a DMR (further broken down into the number of promoters and number of enhancers overlapping a DMR), and the percent of GH elements containing a DMR among all GH elements mapped to the module. DMR = Differentially methylated region. GH = GeneHancer.

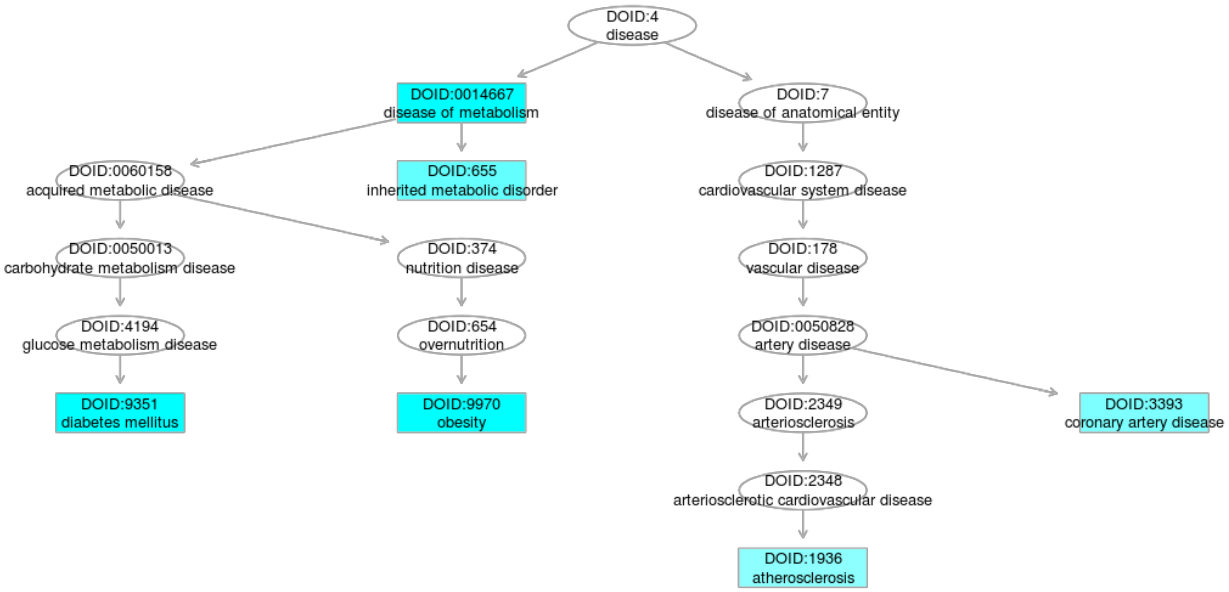
Supplementary Information



Supplementary Figure 1. Directed acyclic graph showing the Disease Ontology structure of the top 15 terms (of 136 with FDR-adjusted p-value < 0.05) from the green module. Terms with significant enrichment are in box-shaped nodes, and darker color indicates a more significant p-value. The full list of diseases enriched for genes in the green module and associated gene names are in Supplementary File 3. The relationships of green module Disease Ontology terms in not pictured can be explored interactively at http://disease-ontology.org/.



Supplementary Figure 2. Directed acyclic graph showing the Disease Ontology structure of all terms with FDR-adjusted p-value < 0.05 (n=11) from the black module. Terms with significant enrichment are in box-shaped nodes, and darker color indicates a more significant p-value. The full list of diseases enriched for genes in the black module and associated gene names are in Supplementary File 3.



Supplementary Figure 3. Directed acyclic graph showing the Disease Ontology structure of all terms with FDR-adjusted p-value < 0.05 (n=6) from the cyan module. Terms with significant enrichment are in box-shaped nodes, and darker color indicates a more significant p-value. The full list of diseases enriched for genes in the cyan module and associated gene names are in Supplementary File 3.

Module	Number of Genes	Total GH elements	GH elements with RELA TFBSs	Percent of total GH elements	GH elements with TFBSs and DMRs	Percent of total GH elements
Green	1,067	29,456	3117	10.6%	74	0.25%
Purple	491	13,144	1263	9.6%	5	0.04%
Black	679	18,924	1816	9.6%	22	0.12%
Brown	1,633	44,637	4001	9.0%	33	0.07%
Turquoise	2,570	55,670	5625	10.1%	50	0.09%
Greenyellow	221	5,538	556	10.0%	3	0.05%
Tan	187	3,918	433	11.1%	2	0.05%
Red	828	20,300	2115	10.4%	15	0.07%
Salmon	122	3,581	327	9.1%	2	0.06%
Yellow	1,252	18,503	1843	10.0%	19	0.10%
Cyan	37	672	52	7.7%	2	0.30%
Midnightblue	34	396	43	10.9%	1	0.25%
Pink	673	12,661	1233	9.7%	16	0.13%
Magenta	587	13,906	1466	10.5%	17	0.12%
Blue	2,202	59,406	5932	10.0%	54	0.09%
Grey/unassigned	1,436	30,967	2772	9.0%	26	0.08%

Supplementary Table 1. Distribution of known HUVEC RELA transcription factor binding sites (TFBSs) across the GH elements of genes in modules.

Supplementary File 1. Names and WGCNA module assignments of 14,019 genes expressed in untreated and TNFa-treated HUVEC cell lines, and measures of connectivity (kTotal, kWithin, kOut, kDiff), gene significance (GS), and module membership (MM). The second sheet shows the MM value of every gene for every module. The third sheet contains the differential expression (DE) test results from limma. (Separate file chapter5_supplementary_file_1.xlsx.)

Supplementary File 2. Differentially methylated regions identified by Bumphunter. Columns are: chromosome, start position, and end position of the region; value = average of the estimated regression coefficient; area = the absolute value of the sum of estimated coefficients for the region; L = the number of probes in the region; clusterL = the number of probes in the cluster (not all probes in the cluster are necessarily included in the region); p.value = p value for differential methylation; fwer = p value for differential methylation corrected to account for the family-wise error rate. The remaining columns identify genes that (1) contain a DMR in the gene body or (2) have 5' untranslated regions (5' UTRs) within or near a GH promoter containing a DMR. The second sheet shows traits associated with SNPs in genes with DMRs in their promoters. (Separate file chapter5_supplementary_file_2.xlsx.)

Supplementary File 3. Complete Disease Ontology results for the green, black, and cyan modules. Columns include Disease Ontology identifier; disease name; number of genes annotated for the disease; number of genes in module that overlap the disease annotated genes; enrichment fold change, z-score, p-value, FDR-adjusted p-value, odds ratio, 95% confidence interval upper and lower bounds; list of annotated genes for the disease, and list of in module that overlap the disease annotated genes. (Separate file chapter5_supplementary_file_3.xlsx.)

Chapter 6 - Conclusions

In this dissertation, I examined DNA methylation in immune cells obtained from blood in two case-control studies of RA and MS, inferred miRNA contributions to pediatric-onset MS from GWAS data, and characterized DNA methylation and gene expression changes in endothelial cells that follow stimulation by the inflammatory cytokine TNFa. This chapter summarizes the key findings of each of these studies.

Chapter two compared DNA methylation of 63 RA cases to 31 healthy controls in four sorted immune cell types collected from blood samples. Approximately 430,000 CpG sites (cytosine bases followed by guanine bases that are potentially methylated) were examined. CpG sites previously identified in fibroblast-like synoviocytes (FLS) obtained from joints of RA cases and found to have increased methylation compared to osteoarthritis cases or healthy controls were found to also be hypermethylated in CD4+ naïve T cells from RA cases relative to healthy controls. These results show a disease-associated signature can be observed in cells obtained from whole blood, which is much more accessible for clinical and epidemiologic studies compared to synovial fluid. FLS-representative DNA methylation signatures derived from blood may prove to be valuable biomarkers for RA risk or disease status.

Chapter three compared DNA methylation in 94 women with MS and 94 healthy women in two immune cell types, CD4+ and CD8+ T cells, also isolated from blood samples. Four regions of markedly increased or decreased methylation in MS cases were found, providing evidence that DNA methylation of CD4+ and CD8+ T cells plays a role in MS etiology. Genes near regions of differential methylation were subsequently tested for differences in gene expression levels in a separate sample of female MS cases and healthy controls. Differentially methylated regions (DMRs) in the *SLFN12* and *HLA-DRB1* genes were consistently observed across the two T cell sub-types, and differential gene expression was detected in whole blood for these gene candidates. Results indicate that DMRs may be detected in more accessible whole blood samples, paving the way for future large-scale studies of DNA methylation in MS. The *SLFN12* findings are particularly compelling and warrant further investigation, as this gene is known to be down-regulated during T cell activation and up-regulated by type I interferons, which are already used to treat MS.

Chapter four utilized results from the largest pediatric-onset MS GWAS study to date, and showed that there are likely miRNA contributions to pediatric-onset MS. Using this approach to infer miRNA involvement from GWAS data is especially helpful in a rare disease setting, where collecting and processing blood or other tissue samples is challenging. Analyses showed that miRNA-target gene signals were enriched in GWAS results and identified 39 candidate biomarker miRNA-target gene pairs. The candidate biomarker target genes included immune and neuronal signaling genes. Further, dysregulation of miRNA binding to target genes was implicated in biological pathways involved in immune signaling. These findings provide evidence that pediatric-onset MS risk is influenced by miRNAs acting on specific immune signaling and other genes.

Several miRNA-target gene pairs and specific tissues were nominated for further study, and further work is needed to determine whether the miRNA-mediated disease processes found are specific to the pediatric MS population. The candidate biomarker miRNA-target gene pairs should be further studied for diagnostic, prognostic, and/or therapeutic utility.

Chapter five characterized DNA methylation and gene expression changes that occur in human endothelial cell lines after activation with the inflammatory cytokine TNFa. Disease Ontology enrichment analysis was performed to better understand the gene expression changes. Weighted gene correlation network analysis (WGCNA) identified 15 gene modules, or sets of genes related by similar expression patterns. Four modules showed a strong association with TNFa treatment, indicating that those sets of genes act in concert in response to increases in TNFa in endothelial cells. Genes in the top TNFa-associated module were all up-regulated, had the highest proportion of hypomethylated DMRs, and were associated with 136 Disease Ontology terms, including infectious, autoimmune, and cardiovascular diseases, and cancers. These results can inform future research on new drug targets for diseases that are currently treated with TNFa inhibitors.

In summary, each of the studies detailed in this dissertation took an epidemiologic approach and employed a variety of computational methods to further the understanding of autoimmune disease pathogenesis and identify avenues of future research into diagnostic, prognostic, and therapeutic tools to combat these diseases.

Dissertation Publications

1. Rheumatoid arthritis naive T cells share hypermethylation sites with synoviocytes. (Chapter 2)

   Brooke Rhead, Calliope Holingue, Michael Cole, Xiaorong Shao, Hong L. Quach, Diana Quach, Khooshbu Shah, Elizabeth Sinclair, John Graf, Thomas Link, Ruby Harrison, Elior Rahmani, Eran Halperin, Wei Wang, Gary S. Firestein, Lisa F. Barcellos, and Lindsey A. Criswell. *Arthritis Rheumatol.* 2017 Mar;69(3):550-559.

2. Increased DNA methylation of SLFN12 in CD4+ and CD8+ T cells from multiple sclerosis patients. (Chapter 3)

   Brooke Rhead, Ina S. Brorson, Tone Berge, Cameron Adams, Hong Quach, Stine Marit Moen, Pål Berg-Hansen, Elisabeth Gulowsen Celius, Dipen P. Sangurdekar, Paola G. Bronson, Rodney A. Lea, Sean Burnard, Vicki E. Maltby, Rodney J. Scott, Jeannette Lechner-Scott, Hanne F. Harbo, Steffan D. Bos, Lisa F. Barcellos. *PLoS One.* 2018 Oct 31;13(10):e0206511.

3. miRNA contributions to pediatric-onset multiple sclerosis inferred from GWAS. (Chapter 4)

   Brooke Rhead, Xiaorong Shao, Jennifer S. Graves, Tanuja Chitnis, Amy T. Waldman, Timothy Lotze, Teri Schreiner, Anita Belman, Lauren Krupp, Benjamin M. Greenberg, Bianca Weinstock–Guttman, Gregory Aaen, Jan M. Tillema, Moses Rodriguez, Janace Hart, Stacy Caillier, Jayne Ness, Yolanda Harris, Jennifer Rubin, Meghan S. Candee, Mark Gorman, Leslie Benson, Soe Mar, Ilana Kahn, John Rose, T. Charles Casper, Hong Quach, Diana Quach, Catherine Schaefer, Emmanuelle Waubant, Lisa F. Barcellos, on behalf of the US Network of Pediatric MS Centers. *Ann Clin Transl Neurol.* 2019 May 15;6(6):1053-1061.

4. TNFa drives DNA methylation and transcriptional changes and activates autoimmune disease genes in endothelial cells. (Chapter 5)

   Brooke Rhead, Xiaorong Shao, Hong Quach, Poonam Ghai, Lisa F. Barcellos, Anne M. Bowcock. *Submitted*.